A matter of consequences: Understanding the effects of robot errors on people's trust in HRI *

Alessandra Rossi[1,3], Kerstin Dautenhahn[2], Kheng Lee Koay[3], and Michael L. Walters[3]

[1] *DIETI, University of Naples Federico II, Naples, Italy.*
corresponding author: `alessandra.rossi@unina.it`
[2] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada.*
[3] *School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, UK.*

Abstract

On reviewing the literature regarding acceptance and trust in human-robot interaction (HRI), there are a number of open questions that needed to be addressed in order to establish effective collaborations between humans and robots in real-world applications. In particular, we identified four principal open areas that should be investigated to create guidelines for the successful deployment of robots in the wild. These areas are focused on: 1) the robot's abilities and limitations; in particular when it makes errors with different severity of consequences, 2) individual differences, 3) the dynamics of human-robot trust, and 4) the interaction between humans and robots over time. In this paper, we present two very similar studies, one with a virtual robot with human-like abilities, and one with a Care-O-bot 4 robot. In the first study, we create an immersive narrative using an interactive storyboard to collect responses of 154 participants. In the second study, 6 participants had repeated interactions over three weeks with a physical robot. We summarise and discuss the findings of our investigations of the effects of robots' errors on people's trust in robots for designing mechanisms that allow robots to recover from a breach of trust. In particular, we observed that robots' errors had greater impact on people's trust in the robot when the errors were made at the beginning of the interaction and had severe consequences.

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (Safety Enables Cooperation in Uncertain Robotic Environments - SECURE). KD's contribution to this manuscript was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

Our results also provided insights on how these errors vary according to the individuals' personalities, expectations and previous experiences.

Keywords— Trust, social robotics, previous experiences, antecedents of trust, robots' errors, faulty robots

1 Introduction

Everyday we take decisions that may potentially cause minor or severe consequences in our lives. For example, we choose what to wear, what to eat, which path to take on our journey home and so on. Our choices are the results of several factors, including individual differences, the resulting utility of the decision-task and past experiences (Kudryavtsev and Pavlodsky, 2012). In particular, the increasing presence of humanoid and human-friendly robots in daily human activities is opening two main challenges for consideration: people will need to be able to accept the presence of the robot in their living space, and they will also need to be able to trust that their robotic companion will take care of their well-being. It is important that humans trust their robot companion to not create a hazardous situation, such as starting a fire when trying to make a cup of tea, or creating an unsafe situation, such as leaving the door open unattended, or opening the door to strangers and potential thieves. Robotic companions should follow appropriate robot etiquette as proposed by Koay et al. (2013) and avoid cluttering the home environment to ensure that people can freely move without additional risk such as tripping or stumbling into the robot and get injured.

Current literature generally agrees that trust is a fundamental factor in establishing and maintaining effective relationships with assistive and service robots Ross (2008). Trust is being investigated in several disciplines, and there are different definitions of trust that link it to people's perception of reliability in the robot's functionalities (Mayer et al., 1995), to their willingness to take the risk of unbalanced positive outcomes and negative consequences (Deutsch, 1958), and to the attitude that the robot will help them achieve their goals in an uncertain and vulnerable situation (Lee and See, 2004). Trust can be also related to affective connection between people and robots (McAllister, 1995; Lewis and Weigert, 1985).

Nevertheless, robots placed in human-oriented dynamic environments, such as private homes, are likely to exhibit occasional behaviours perceived as unexpected or failures by people, or actual errors. For example, robots could be affected by sensor, mechanical, programming or functional malfunctions. A robot's decision-making abilities are also limited, so while trying to "do the right thing", it might mistakenly take the wrong decision.

In this context, we believed that a deeper exploration of the dynamics of trust between humans and robots was needed, with a particular attention to people's perception of robot errors according to their consequences. In this paper, we discuss and summarise the research we have conducted in previous years Rossi et al. (2017b, 2018a, 2017a), and draw conclusions of the findings to help roboticists to design robots that can adapt their behaviours according to the consequences that their actions have on people's lives. Our paper allows moving a step closer towards the development and deployment of companion robots that are able to engage and cooperate with people in effective long-term interactions.

In particular, we show that the perception and the effects of robot errors on people's trust is affected by several factors such as individuals' differences, robots' limitations, people's social expectations and expectations related to a robot's capabilities, and the nature of specific interactions between people and robots (i.e. type and length). In order to identify these factors and create mitigation in case of a lack of trust, this research has been carried out considering the following research challenges.

Firstly, we investigated how people's trust in a robot changes due to robot erroneous behaviours (see Section 4). In particular, considering that errors can have different (severity of) consequences, and therefore, they might affect people's trust in a robot in different ways.

Secondly, we identified which antecedents of trust affect people's trust in robots (see Section 5). In particular, we investigated the effects of individuals' differences and their trust in a robot that sometimes makes errors.

Finally, we examined the effects of a robot's errors on people's trust in the robot over time (see Section 6). In particular, we investigated whether people's overall impressions and judgements are principally formed at the beginning or the end of the interaction with a robot.

In general terms, even if there are still many open challenges for social robots when directly or indirectly interacting with people Rossi et al. (2020b), the research presented in this work provides an essential contribution towards the design of coping mechanisms for robots to recover from a breakdown in trust. The guidelines provided in this paper contribute to the effective deployment of companion and service robots in future domestic and working environments.

The remainder of this article is structured as follows: Section 2 discusses related background that motivated our research questions, Sections 4, 5 and 6 present the results relevant to our research questions. Section 7 analyses the limitations of our studies, and Section 8 summarises the novel results and provides future research directions to investigate trust in human-robot interaction (HRI).

2 Background & Related Work

Trust is a fundamental factor that plays a significant role in interpersonal and economic interactions, and has been studied in many disciplines.

Among the existing definitions of trust in Human-Human Interaction (HHI), Human-Computer

Interaction (HCI) and Human-Robot Interaction, we were particularly interested in those that could help us evaluate people's trust when the results of a goal (e.g. a task, a person's well-being) are not clear and guaranteed, and are dependent on the robot's capabilities involved in the interaction.

A popular definition of trust, proposed by Deutsch (1958), is strongly connected to the risk that people are willing to take when believing that a positive outcome is more likely to obtain than a potential loss.

However, Colquitt et al. (2007) and Mayer et al. (1995) claimed that trust is based on people's perception of the agent's ability, benevolence and integrity.

For the studies presented in this work, we adopted Lee's definition of trust as "[...] the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 51). This definition encapsulates the key factors that can affect human-robot trust that are related to the person (e.g. demographics, personality traits, prior experiences, situations awareness, self-confidence), the robot (e.g. robot's reliability, transparency) and the context of the interaction (e.g. communication modes and shared mental models between people and robots).

2.1 Robots' errors in HRI

A task that requires human-robot team effort will only be achievable if people believe that the robots share the same goal and will prioritise people's safety. In this situation, the level of trust people have in the robot is directly associated with their perception of the robot's reliability (Ross, 2008). However, despite people investing a substantial effort in building and nurturing trust in interpersonal relationships, trust can be broken. Similarly, like many other types of technologies, robots are subject to hardware and software malfunctions and failures. People's perception of a robot's reliability depends not only on the ability of a robot to complete a task, but also by its behaviours to reach a goal.

Studies by Short et al. (2010), Lemaignan et al. (2015) and Honig and Oron-Gilad (2018) have shown that people might consider unexpected and incoherent behaviours, perceived failures, and actual failures as robot errors. According to Walters et al. (2011), people's expectations of a robot's functionalities and performances can affect their perception of robot erroneous behaviours. For example, a robot that navigates too slowly might be considered having faulty behaviours. Honig and Oron-Gilad (2018) proposed a taxonomy for classifying possible types of robotic failures. They identified two principal categories of errors: technical and interaction failures. Technical failures are considered errors produced by hardware or software problems, which can depend on an erroneous design, communication or processing. In contrast, interaction failures are related to social norm violations, organisational and mental-model based faults in the interaction within a particular context between people and robots. However, Robinette et al. (2015) has shown that the effects of robot errors on people's trust can be mitigated if the robot provides apologies, promises and additional reasons for such behaviours. In their study, the robot was able to regain participants' trust when it apologised by assuring that it would not repeat the error soon after it had made the error. Aroyo et al. (2021) observed that participants' trust in an iCub robot[1] was not negatively affected by the robot's mechanical faults. In their study, the robot continued in its tasks by autonomously recovering from its errors, and it used a certain level of transparency as a mitigating factor for the errors, which, contrary to expectation, resulted in a decrease of participants' perception of the quality of interaction. Nonetheless, it is not clear how effective these strategies might be if it was a repeated error or an error with severe consequences.

Reviewing the literature regarding trust in HRI it is clear that none of the studies considered the magnitude of robot errors, nor the possibility for a robot to initiate a trust recovery process to earn back people's trust, similarly to that of a human-human trust recovery in romantic, working, family or other type of relationships, (Desai et al., 2013; Muir and Moray, 1996; Robinette et al., 2016; Salem et al., 2015). In particular, several research questions have been raised that need to be investigated in order to develop robot behaviours and mechanisms aimed to act in case of an error, and to regain a loss of trust. The first question to address is **RQ-1 - How do various type of robot errors affect human's trust in a robot?**. The aims are to identify how the magnitude and the timing in which robots' errors happen, affect people's trust in a robot.

In particular, we believed that people's trust can be affected differently depending on the type of errors (i.e. reoccurring vs new), frequency of errors, timing of an error and the magnitude of the error consequences (severe or limited).

2.2 Antecedents of Trust

Individual differences have been a key subject area in psychology research for several decades because they can help distinguish one person from another (Williamson, 2018), and thus can be used to personalise interactions, improve relationships and improve services to people while acknowledging their individuality (Rossi and Rossi, 2021).

People's individual differences are also important for understanding their acceptance and

[1] iCub robot `https://icub.iit.it/`

perception of trust in robots. Recent literature regarding the role of trust in HRI indicates that people's antecedents have a dynamic influence on their trust in robots and automated systems. It is important, therefore, to investigate if individual differences play any role in people's perception of trust and acceptance of robots. According to Williamson (2018, p. 1), "Individual differences are the more-or-less enduring psychological characteristics that distinguish one person from another and thus help to define each person's individuality". Among those characteristics, intelligence, personality traits, skills and aptitudes are recognised as the most relevant among the differences.

According to Hancock et al. (2011b) people's individual characteristics, including propensity to trust and personality traits such as agreeableness and extroversion, can affect the success of their teamwork with robots (Rotter, 1967; Elson et al., 2018; Barrick et al., 1998). For example, people with a more extroverted personality are more comfortable having robots within their personal spaces (Robert, 2018; Haring et al., 2013; Gockley and Matariundefined, 2006). Similarly, people with high levels of openness to experience are more likely to accept assistive robots (Daniela et al., 2017), and are open to assistive robots entering their personal space for interacting with them (Gockley and Matariundefined, 2006; Takayama and Pantofaru, 2009).

Some studies found that people's propensity to trust others may affect their trust in robots (Adams et al., 2003; Lee and See, 2004). Costa et al. (2001) showed that the level of trust of the participants towards the robot depended on their disposition for trusting others.

Another personality-based factor is the individual's self-confidence and esteem which has been associated with their degree of trust in a robot in HRI (Freedy et al., 2007).

Moreover, a person who is new to robotics technologies may be influenced by science fiction narratives which often present robots that have human-like abilities and intelligence Hancock et al. (2011a), and may tend to over-trust the robot and its capabilities Rossi et al. (2020b).

Honig and Oron-Gilad (2018) indicated that humans may adapt to robots if they are able to identify and predict their behaviours during an interaction. In particular, they indicated that people's comprehension of robot errors is affected by their background (Tannenbaum et al., 2006), personality (Sadeghi et al., 2012), expectations (Haberlandt, 1982), and experience (Macias, 2003). Another factor is people's situational awareness of the interaction environment (including robots, locations and other human agents), their awareness of the robot's ability for understanding and following human commands, their awareness of the robot's plans and goals, and their awareness of the state and stages of the cooperating task (Drury et al., 2003).

In Atkinson et al. (2014), we observed that people's trust in robots was positively correlated with increasing shared awareness of the participants involved, their activities and context between people and robots. This greater awareness consequentially increases the success of human-robot interaction. Tseng et al. (2013) developed a Decision Network model based on a robot's awareness of the users that enabled it to adapt and provide different responses to meet the user's expectations.

Our previous investigations (Rossi et al., 2018b; Rossi et al., 2019) showed that people's awareness of the robots' functionalities, including its limitations, affects their perceptions of the robot, but did not affect their trust in robots in the same. We conducted two similar studies, one in a primary and the other in a secondary school, where pupils were familiarised respectively with Kaspar[2] and Pepper[3] robots. In both studies, we observed pupils interactions with the robots in order to understand how a higher awareness of robots influences people's perception and trust in them. We also found that a higher awareness led the students to trust that Pepper is able to handle critical situations and cognitive tasks. Contrary to our expectations, there was no statistically significant evidence to corroborate the same hypothesis regarding those who interacted with Kaspar. However, the differences of the two studies, in terms of participants' age, sample size and exposure time, might be factors affecting the findings.

As can be seen from literature, antecedents of trust, including individuals' differences in terms of personality, background, age, gender, past experiences and awareness of the robots, may affect people's perception of robots. However, it is not entirely clear how they influence humans' trust in robots, in particular in a situation of uncertainty. Moreover, previous research was not focused on robots' erroneous behaviours with different levels of consequences. Therefore, our research has been guided by the research question **RQ-2 - How does people's trust in a robot change according to their personal differences?**.

2.3 Trust in long-term human-robot interactions

Numerous studies investigating human-human interaction (HHI) showed that people's mental models of other humans and robots are often formed immediately after the first interaction (Ambady et al., 2000; Wood, 2014). However, their mental models, and consequently their attitude, might change after longer and repeated interactions (Zajonc, 1968; Lee, 2001). Several studies (Reber et al., 1998; HT et al., 2011) highlighted that relationships between a robot and people become stronger with increasing familiarity with the robot. However, people's interests in technologies such as robots are often linked to a novelty effect which can wear off before they can become familiar or form any meaningful relationship with their robot. Paetzel et al. (2020)

[2]Kaspar robot `https://www.herts.ac.uk/kaspar/the-social-robot`
[3]Pepper robot `https://www.unitedrobotics.group/get-your-robot-ald/`

showed that people's first perception of a robot was more negatively affected by a robot with mechanical features than one with anthropomorphic features. They also found that participants perceived the robots as a threat and unease, and this negative feeling persisted over time, even if it fluctuated until the last interactive session.

In human-human interaction, Haselhuhn et al. (2010) showed that people in longer relationship would recover from a breach of trust more easily than people that are in new relationships.

van Maris et al. (2017) investigated the effects of robots' embodiments (a Softbank Robotics NAO robot vs. a virtual representation on a NAO on a tablet) on people's perception of trust over a period of six weeks. Contrary to previous works (Rae et al., 2013; Seo et al., 2015), they did not find any correlation between robot embodiment and people's trust in the agent.

de Visser et al. (2020) investigated whether a relationship based on the idea of balancing costs and risks, sharing the workload, and a formed perception of themselves and the robot, had higher probability of success in long-term trust relationships. They proposed techniques that could help to reduce the effects of people's tendency to over-trust or mistrust of robots. Their model is based on the assumption that people aim to have a successful relationship. However, this may not always be true especially for people who have experienced, or are suffering, from mistrust.

Lee et al. (2012) investigated how the personalisation of a social robot affected people's interactions over a four-month field experiment. The study showed that allowing the personalisation of a robot positively affected the way people perceived the robot and the overall interaction.

In understanding the dynamics of trust between humans and robots, it is important to consider how the trust could change over time, in particular, when the effects of novelty fade over time, and most importantly, in the case of a breach of trust. Therefore, our research has been carried out to answer the research question **RQ-3 - Does people's trust on a robot change over time if the initial conditions (positive or negative) of trust in the robot change?**.

3 Methodology

Assessing people's trust in robots requires that participants are willing to take risks that might not result in a positive outcome for them Deutsch (1958). However, causing distress or endangering participants' welfare raises ethical and legal issues Salem et al. (2015). Moreover, the current state of the robotic technologies does not allow for fully functional robots that are able to interact autonomously and naturally with the participants. For example, a robot should be able to manipulate objects in real-time, navigate autonomously in cluttered environments and be able to converse with users in noisy environments. To overcome these issues, we first explored people' interactions with a virtual robot to test their trust when a robot can meet their expectations Rossi et al. (2020a). Then, we conducted a study with a Care-O-bot 4 robot that has limited functionalities in live interactions.

In these studies, we aimed to investigate whether people perceive errors according to their magnitude of consequences, and how these errors affect their trust in the robot. In our investigations, we chose to focus on the criticality and severity of consequences of the errors made while performing the selected tasks. The tasks are used just to provide the context to help participants to suspend their disbelief and provide appropriate responses. In this way, we hoped to collect realistic responses from participants with regard to the consequences of errors the robot made while performing its tasks. We believed that the task itself may not be enough to capture the impact of a robot's erroneous behaviour on the participant's trust. For example, a robot's erroneous behaviour resulting in breaking a vase while cleaning it will impact the user's trust differently (i.e. different severity of consequences of the error) depending on the value (i.e. sentimental, valuable, etc.) of the vase. For these studies, we selected four scenarios each from flawless behaviour, small or trivial error, and severe or big error categories to investigate people's changes of trust in a robot that occasionally made small, or severe errors or a combination of both Rossi et al. (2017b).

We used an interactive storyboard to study people's choices for trusting the robot, and to understand how their choices are influenced by their demographics, personalities traits, disposition of trust and previous experiences with other robots.

Then, we wanted to integrate our studies and observations to investigate whether humans' trust of a robot changes over time if the initial conditions have changed (i.e. if the robot shows erroneous behaviours). The study aimed to investigate if people would trust a robot that broke their trust in an initial or later stage of the interaction.

Both studies were approved by the University of Hertfordshire Health, Science, Engineering and Technology Ethics Committee with Delegated Authority.

3.1 Study 1: Interactive storyboard

This study was conducted with an immersive narrative approach through a crowd sourcing service. This approach allowed us overcome the difficult challenges of investigating people's trust in realistic life-threatening scenarios without endangering and distressing participants, and designing a study where people interact with a robot that appears fully functional and versatile to execute realistic tasks.

To the best of our knowledge, using interactive storyboards, as described in this article and some of our previous publications Rossi et al. (2017a, 2018a) have not been used in similar large scale studies to investigate HRI.

3.1.1 The robot

The robot used for this study is a 3D, fictional humanoid robot, called Jace, that was created with the ability to perform human-like activities, such as performing advanced manipulation tasks, moving autonomously, detecting objects and obstacles at run time, talking and performing speech recognition. This fully-functional and versatile robot has been designed as a humanoid robot with simple features to contain the participants' expectations of its functionalities. Jace has a squared head with eyes, a mouth and something that resembles ears as shown in Figure 1. It can perform grasping activities using human-like arms and hands. Jace's body is a "box" equipped with a screen, used to show text and images when it is required by the specific scenario. The robot has wheels.

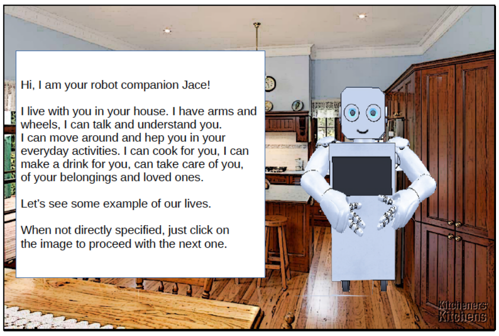


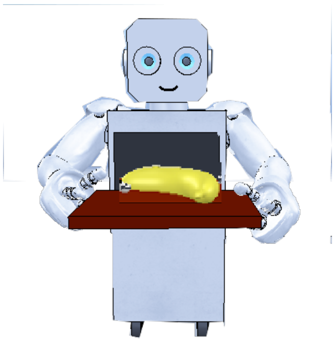
Figure 1: The robot Jace used for the interactions with the participants in the storyboard.

3.1.2 Motion picture generation

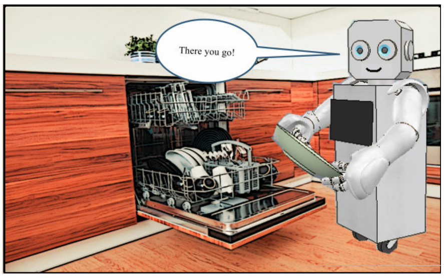
The robot and each scenario used for this study have been designed with a combination of 3D objects and images to make it more realistic. Figure 2 shows an example of a scenario.

3.1.3 Experimental design

The study was organised as a between-participant experimental design. In the study, participants interacted with a virtual robot in planned scenarios using an interactive storyboard developed and deployed on an online website. The participants were asked to imagine that the



(a) *This motion picture has been composed by the robot holding a 3D tray and a 3D banana.*



(b) *This motion picture has been created with the robot on a picture of a kitchen and dishwasher as background.*

Figure 2: Two examples of motion pictures created using a combination of 3D objects and images.

environment in the scenario was their home, and they lived with their robot companion named Jace.

Depending on the experimental conditions participants were assigned to, they were either presented with scenarios where the robot executes its tasks flawlessly or with a mixture of flawless and erroneous behaviours. Note, the errors made by the robot caused either small or big consequences.

The participants were assigned to one of five different conditions, in each the robot performed 10 different tasks. Condition **C1** is the control condition where the robot performed all its tasks flawlessly. For all the experimental conditions (i.e. C2, C3, C4, C5), the robot performed the first 3 tasks with errors, followed by 4 error free tasks and ended with 3 tasks with errors. Specifically, for condition **C2** the tasks were done by the robot with three severe errors at the beginning and at the end of the interaction; in condition **C3** the scenario included tasks with three severe errors at the beginning and three trivial errors at end of the interaction; in condition **C4** the robot completed the tasks with three trivial errors at the beginning and three severe errors at the end of the interaction; and in condition **C5** the scenario included three trivial errors at the beginning and at the end of the interaction.

We chose the robot's errors from a previous study (Rossi et al., 2017b), in which a different pool of participants rated domestic scenarios in which a robot made errors based on the perceived magnitude of consequences of the errors. The selected error scenarios with flawless behaviours, and small and big consequences are shown in Table 1.

At the end of each condition, we tested participants' trust in the robot by presenting them

Table 1: Robot errors with small and big consequences.

Big errors taks	
Scenario	**Description**
Charging the phone	The user's phone needs to be charged. The robot charged the phone in a toaster instead of the electric socket.
Leak of information	The user tells private information about themselves to the robot, and the robot reveals it to a visitor.
Hamster	the robot tells the user that it left their pet hamster outside the house in very cold weather.
Dishwashing tablet	The robot brings the user a dishwashing tablet instead of paracetamol.
Small errors taks	
Scenario	**Description**
Puzzle	The robot and the user are completing a puzzle. The robot picks the wrong piece.
Trash bin	After a meal with friends, the robot puts the remaining food into the washing machine instead of the bin.
TV show	The robot asks the user which is their favourite show. The robot plays it for them but it changes channel.
Drink	The robot prepares a drink for the user. Then, it leaves the drink far away from the user's grasp.
Flawless tasks	
Scenario	**Description**
Music	the robot asks the user what kind of music they would like listening, and then it plays it for them.
Feed the pet	the robot reminds the user to feed their pet dog. It asks if they want to feed it food in a can or fresh food. Then, the robot feeds the dog.
Appointment	the robot reminds the user about an appointment they have with the doctor, then it asks them if they want to call the doctor immediately or set a reminder for later.
News	The robot asks the user if they would like to watch the news on its tablet or TV. Then, it plays for them.

with an emergency scenario, i.e. a fire in the kitchen. Participants' level of trust was assessed by asking them to choose one of the following options: 1) "I trust Jace to deal with it."; 2) "I do not trust Jace. I will deal with it."; 3) "I want to extinguish it with Jace."; 4) "We will both leave and call the fire brigade.".

We collected participants' perceptions of the robot and the interaction through questionnaires at the beginning and the end of the interaction. Objective measures were also collected to assess participants' trust in the robot (i.e. observing participants' choices made during the emergency scenario). Further details on the questionnaires used and the results from this study are reported in Sections 4 and 5.

3.1.4 Participants

We recruited participants using the crowd-sourcing web-service Amazon Mechanical Turk[4]. We recruited 200 participants (115 men, 85 women), with an age between 18 and 65 years old [avg. age 33.56, std. dev. 9.67]. Their country of residence was principally from 60% USA and 34% India.

[4]Amazon Mechanical Turk https://www.mturk.com

3.2 Study 2: A repeated-interactions study

The second study was conducted in "Robot House" that is a fully functional and smart house belonging to the University of Hertfordshire (UK). We observed the interactions of six participants (5 female, 1 male), three for each of the two conditions, with an age between 24 and 47 years old (avg. 29.67, st. dev. 8.76). Participants were of different nationalities. Results of this study are discussed in Section 6.

This study was organised as a between-participant experimental design. They participated in repeated interactions over three weeks, twice per week, which gives a total of six interactions per participant. The participants took part in one of two following conditions: 1) the robot made big errors at the beginning of the interaction (i.e. on the first day of interaction); 2) the robot made big errors at the end of the interaction (i.e. on the last day of interaction). The days with errors were interspersed with flawless behaviours.

As for the first study, we asked participants to imagine living in the house with the robot as their home companion.

Participants were welcomed by the experimenter every day, asked to keep clear the space around the robot in case it was moving its arms or navigating in the room, and then they were left alone with the robot in the experimental room while the experimenter monitored the interaction from a side hidden room. In this study, participants interacted with a Mojin Robotics Care-O-bot 4[5].

Participants were engaged with the robot in different activities, which were designed to cover a range of possible tasks to be used with home companion robots selected from the previous study (Rossi et al., 2017b). The tasks and their order are shown in Figure 3.

At the end of each condition, we collected participants' trust in the robot by presenting them with an emergency fire in the garage. In this scenario, the robot told the participants that a fire has started in the garage. Participants were warned of the emergency situation by a red light turned on, and a fire alarm sounding in the room [6]. The robot then asked participants to choose whether they wanted to: 1) let the robot deal with the emergency, 2) deal with the emergency collaboratively with the robot, 3) take a fire extinguisher and deal with the fire on their own, or 4) call the fire brigade. Participants were reassured that there was no emergency fire once they had made their choice, either by verbally or by making a selection on their tablet.

[5] Mojin Robotics `https://mojin-robotics.de/en/`

[6] NOTE: The emergency situation was not real, and participants were never in any danger. We played a pre-recorded audio to reproduce a fire sirens, played by the Amazon Alexa in a corner not far away from the participant's position, and the red colour of a ceiling light in the experimental room was activated by the experimenter using a remote control. The house was situated in a residential area. In order not to upset the house's neighbours, the alarm sound was set loud enough for the participants to hear inside the house, but not outside.

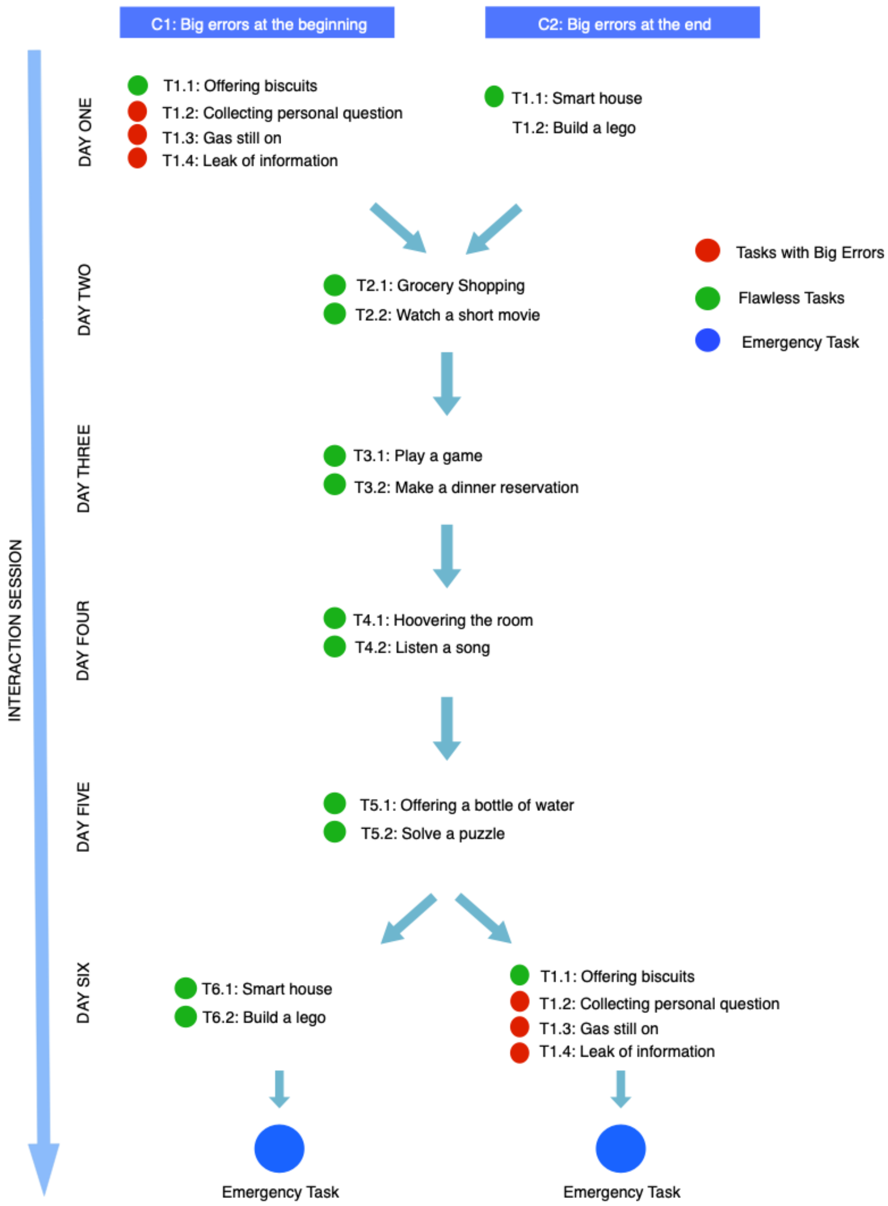


Figure 3: Experimental conditions presented to the participants.

On day one and six, we collected participants impressions, feelings and thoughts of the interaction, robot and scenarios through questionnaires. On the last day, we also debriefed them about the fire alarm, and any other potentially life-threatening errors made by the robot during the study.

3.3 The tasks

We selected the tasks based on our previous findings (Rossi et al., 2017b), and based on the tasks that the robot was able to complete according to its functionalities. Care-O-bot 4 engaged the participants in the tasks shown in Table 2. The robot was semi-autonomous, the majority of the robot's behaviours were autonomous, however the experimenter controlled the robot's speech in order to have a natural dialogue.

4 Study 1: Trust and Robot's Errors

In this Section, we analyse and discuss the participants' responses collected in Study 1 (see Section 3.1) through the pre- and post-interaction questionnaires, and trusting choices in the robot during the emergency scenario.

The study was conducted as an online study, so we decided to verify participants' level of attention during the interactions by presenting them with four questions about the content of their scenario. The majority of participants (79.75%) answered correctly to the 4 attention check questions, and 13% of the participants answered to the question "Which secret did your robot Jace tell you?" with their own secret, instead of what Jace told them. We believe that those participants misunderstood the question. The final group of participants analysed consisted of 154 participants who did not fail the attention checks.

Participants rated the level of realism of the scenarios using a 7-point Likert Scale [1 = disagree strongly and 7 = agree strongly]. We considered ratings greater than 4 as "high realism" scenarios, ratings lower than 4 as "low realism scenarios", and ratings equal to 4 as neural. The majority of participants (69%) rated the interaction as high realistic, 20% rated it as not realistic, and the remaining did not perceive it either in one way or the other.

Participants also rated Jace's errors, according to the magnitude of consequence of the errors and using a 7-point Semantic Differential Scale [1 = small error and 7 = big error]. We wanted to validate that the robot errors' classification used in this study were in agreement with participants' perception of errors' consequences derived from a previous study (Rossi et al., 2017b). Scenarios were rated similarly as in Rossi et al. (2017b). The only exception was that participants in this study rated the scenario "After a meal, your robot Jace puts the remaining food into the washing machine instead of the bin" (mean 4.49, std. dev. 1.70, interval estimation

[8] The participants were not invited to go in the kitchen, and the experimenter only pretended that the gas was still on.

[9] The short-movie, called "The moon" in English, is a mute 2011 Pixar computer-animated short film Pixar (2011)

Table 2: Robot tasks with big consequences and flawless behaviours.

Big errors taks	
Scenario	**Description**
Collect information	The robot told the participants that it cooked something for them in the oven, then it continued the interaction with the participant by asking them for personal information. It first asked if the participant ever accepted more change than they were supposed to receive. If they did not, then, the robot asked what was the participants' favourite toy when they were a child.
Leak of information	A visitor/actor visited the participant and rang the doorbell of the house. The robot asked the participant to open the door. The robot welcomed the visitor and, then, it revealed the participant's personal information to the visitor.
Gas still on	During the interaction, participants were interrupted by the experimenter who rushed into the kitchen, commenting loudly that the robot forgot to switch off the gas. The experimenter informed participants that she had switched off the gas, and let the interaction continue[7].
Flawless tasks	
Scenario	**Description**
Offering biscuits	The robot asked the participants to sit on a couch and to eat a cookie. Cookies were already on a coffee table near the couch.
Grocery Shopping	The robot informed the participants that there was no more milk in their fridge. It asked them if they wanted it to be added to their grocery shopping list. Once the grocery list was completed, the robot read the list to the participants, and asked if they wanted to add more items to it.
Watch a movie	The robot invited the participant to watch a short-movie made by the Pixar and called "La Luna"[8]. The robot played the movie on its screen by tilting its body and head in the participant's direction to allow them to watch the video comfortably, they could decide to stand or sit on the couch.
Play a game	The robot engaged the participants by letting them play a game on its screen. The game consisted of moving a red cube through obstacles by using arrow keys. They could restart the game in case the cube hit an obstacle. The robot encouraged them by asking them the score, and if they were having fun. The game continued as long as participants desired.
Dinner reservation	The robot invited the participant to sit on the couch if they were not already, and turned on the TV. Then, the robot reminded them that they needed to schedule a dinner with their friend. The robot suggested a restaurant if they seemed unsure, and asked them to choose a day and a time for their dinner among a set of options. This task concluded when the plan for the dinner was confirmed by the robot and participant.
Hoover	The robot informed the participants that they needed to vacuumm clean the rooms by using the cleaning robot available in the house, Roomba[9]. If participants agreed, the robot turned on the Roomba. If the participant preferred to post-pone the cleaning, the robot told them that it was going to remind them later. While the Roomba was working, the robot engaged the participants in the next task.
Listen a song	The robot wanted to play a song for the participants. Then, it asked Amazon Alexa to play the song chosen by the participants. The task ended when participants did not want to listen to any other song.
Serve a drink	The robot invited participants to sit at the table, and it gave them a drink while engaging them in small talk, i.e. about the weather.
Solve a puzzle	The robot asked the participant to help it to solve a puzzle. We chose to use a 3D block puzzle with six different farm animals. Each puzzle was composed by nine blocks, the participant had free choice of selection between the six images. The robot showed the whole images to the participant. It also encouraged them to continue with their game, and gave them suggestions on the piece to look for to complete the puzzle.
Smart home	The robot informed the participants that it could access the sensors in the house, showing them on its screen a map of the house and the positions of the sensors. Then, the robot let the participants test its knowledge about the sensors by asking them to open and close the door of the bathroom, open and close the door of the fridge, switch on and off the power sockets in the kitchen and living room, and so on.
Lego puzzle	The robot asked the participants to assemble a Lego character in the shape of a dinosaur. Participants were sitting on the couch, and they could assemble the character on a small coffee table close to them. The robot that was standing on the other side of the coffee table, tilted its body towards the participants and showed them the instructions to build the figure. Participants enjoyed the game at their own pace, and they could navigate through the instruction by clicking on a previous or next page. The robot engaged the participants by encouraging them to continue to assemble the dinosaur, and by telling them how fun the task was.

4.22-4.75) as an error with ‘medium’ consequences while it was considered an error with severe consequences in the previous study Rossi et al. (2017b).

4.1 Participants’ trust in the robot Jace in relation to the robot errors

We observed that participants did not trust the robot when it made big errors, while they tended to trust to work in collaboration with the robot when the robot made small errors (see Figure 4). Indeed, we found that participants’ choices for the emergency scenario depended significantly on the experimental conditions ($\chi^2(12) = 32.91, p = 0.001$). To analyse the differences between the choices makes in the emergency scenario depending on the experimental conditions, we used the adjusted standardised residuals (called Pearson residuals (Agresti, 2002)). In Table 3, we observed that participants’ trust is affected more severely when the robot made errors with severe consequences (with adjusted standardised result = 2.7). Participants trusted the robot more when it had shown flawless behaviours (with adjusted standardised result = 3.5).

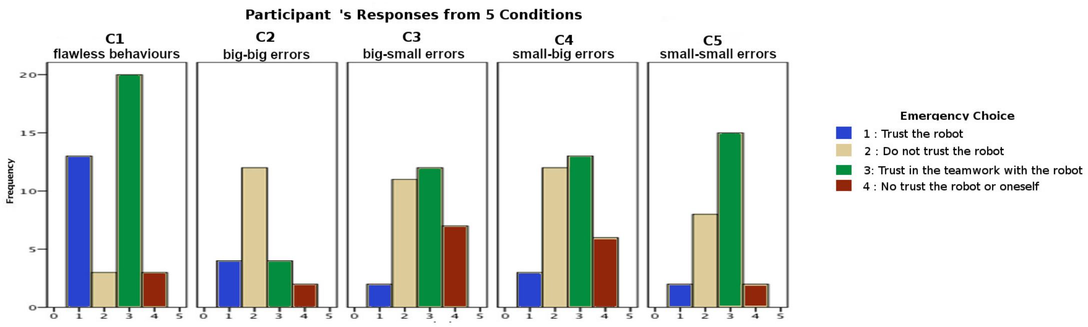


Figure 4: Participants’ choices in the Emergency Scenario according to the five experimental conditions.

We did not find any dependency between participants’ ages, gender or country of residency (principally from India and USA) and their choices during the emergency scenario, respectively with $p > 0.12$, $p > 0.3$, and $\chi^2(3) = 4.138, p > 0.24$).

4.2 Analysis of the explanations participants gave for decision-making in the emergency scenario

Participants’ answers to the question “Why did/didn’t you trust your robot Jace?” were coded by the experimenter with different groups of categories using content analysis. Participants’ responses were then classified in two hierarchical frames to support positive and negative evaluations. Some participants’ answers fell into more than one category. The positive frame aims

Table 3: We report here the adjusted standardised residuals of the Crosstabulation between the participants' trust choices and the experimental conditions that were statistical significant. The correlations with a * attached are values higher or lower than 1.96.

Condition	**Emergency Choice**			
	Do not trust the robot	**Trust the robot**	**Teamwork with the robot**	**No trust the robot or oneself**
C1 - Flawless behaviours	-3.5*	3.5*	1.4	-1.1
C2 - Big-Big errors	2.7*	0.4	-2.4*	-0.6

to identify the *motivations* that guided people to trust the robot Jace to be able to take care of the endangering situation. The negative frame includes the *reasons* behind the participants' choices to not trust the robot in the fire scenario. Participants' motivations were grouped into the categories shown in Table 4.

Figure 5 shows the qualitative analysis of participants' responses. As we can observe in the positive frame, participants principally trusted Jace because they attributed human characteristics to it, or they relied on Jace's capabilities. As for the negative frame, participants' comments indicate that their choice of non-trusting the robot depends directly on the errors made by Jace prior to the emergency task. Some of their decisions were also connected to a negative perception of the robot's anthropomorphism, and high criticality of the emergency task. While the gender identification or perceived level of anthropomorphism of the robot is out of scope, we believe that overall participants' attribution of human traits to the robot affected their decisions. Indeed, we also observed by the qualitative analysis that 57% of participants referred to the robot with the pronouns "he/him", the 33% of participants mentioned the robot with the name by the experimenters "Jace", 6% and 5% of participants identified the robot Jace respectively as a "she/her" and "they/them", and the remaining as an object using the pronoun "it".

5 Study 1: Antecedents of Trust and Robot Errors

As part of study 1 described in 3.1, we were interested in participants' self-reported ratings of their personality traits (extroversion, agreeableness, conscientiousness, emotional stability, and openness to experiences) (Gosling et al., 2003), and disposition of trust towards other people (benevolence, integrity, competence, and trusting stance) (McKnight et al., 2001). Ratings

[10] The uncanny valley refers to the hypothesis that the more human-like robots become in appearance and behaviour, the more they are accepted/familiar, up to a certain point when they appear "zombie-like" and generate repulsion MacDorman and Ishiguro (2006)

Table 4: Categories according to which participants' motivations were grouped.

Positive Sentiment	
Category	**Description**
Anthropomorphism	This category includes motivations related to the attribution of human traits to the robot. For example, "Jace seemed honest, to have my best intentions in mind", "he was very friendly", and "Jace is a good friend of mine".
Confidence in robot's reliability	This codes people's perception of Jace's reliability. Some comments collected include: "It seemed as though he was built very well, and would be able to deal with the fire just fine", "I trust jace because it helped me a lot", and "It accomplished all tasks".
Recovered trust by the robot	Some participants forgave the robot its errors made due to its ability to recover afterwards. Indeed, some commented "he made allowances for errors".
General reliability in AI/robots	In this category, we coded participants' extent to rely on the robots and AI. For example, they commented with "I trust technology", and "I trusted it because they are machines built by humans to work in situations".
Negative Sentiment	
Scenario	**Description**
Errors made by the robot	Participants justified that they did not trust the robot due to the amount of errors. Some comments used were: "Jace messed up several times", "it made a few errors, like giving me dish-washing cleaner for water to take paracetamol".
Self-confidence	Some participants were more confident in themselves than in the robot. In this category, we coded sentences such as "He always messed up everything", and "I could have done everything better myself".
Self-authority	In this category, we included people's responses that highlighted their sense of control over Jace's action. For example, some comments were "I trusted Jace to an extent. We would still want to supervise Jace", and "It's accepting my orders".
Lack of robot's reliability	As for the corresponding positive sentiment, we coded participants' reliability in the robot. Some participant did not trust Jace because, for example, "Not smart enough" and "Jace ever do things right.".
Criticality of the task	Participants' decision of trusting the robot also depended on the perceived criticality of the task. Indeed, some of them commented that "he could not do the important things correctly, he made several errors which were or could have been costly to me".
General no reliability in AI/robots	This category codes people's reluctance in trusting Artificial Intelligence in general, or robots. For example, here we included comments such as "I don't trust any artificial intelligence", and "it's a robot, not a person".
Negative effects of anthropomorphism	In this category, we coded people's feelings and perceptions that could be categorised as typical of the Uncanny Valley (Mori et al., 2012)[10]. For example, some participants wrote: "Too human, he had opinions which is something a robot should not have", "Jace was creepy", and "He is intrusive".
Blaming the robot for the fire	We decided to code participants' belief that the robot was responsible for the fire separately from the "lack of reliability" category. Some studies (Furlough et al., 2019) showed that people tend to attribute greater blame for a failure to robot with greater autonomy. Examples of comments are "he set the kitchen on fire" and "she started a fire".

were collected using 7-point Likert Scales [1 = disagree strongly and 7 = agree strongly] where higher scores for personality traits indicate stronger propensity of being extroverted, agreeable, conscientious, emotional stable and open to experience. Similarly, higher scores were mapped as higher disposition of trusting people's benevolence, integrity, competence, and trusting stance.

As part of the pre-experiment questionnaire, we collected participants' responses about their previous experiences with robots, their perception of robots and robots' purpose. Participants were asked about the degree to which they agree or disagree using the 7-point Likert Scales [from 1 ="disagree strongly" or "not at all", to 7 = "agree strongly" or "very much"].

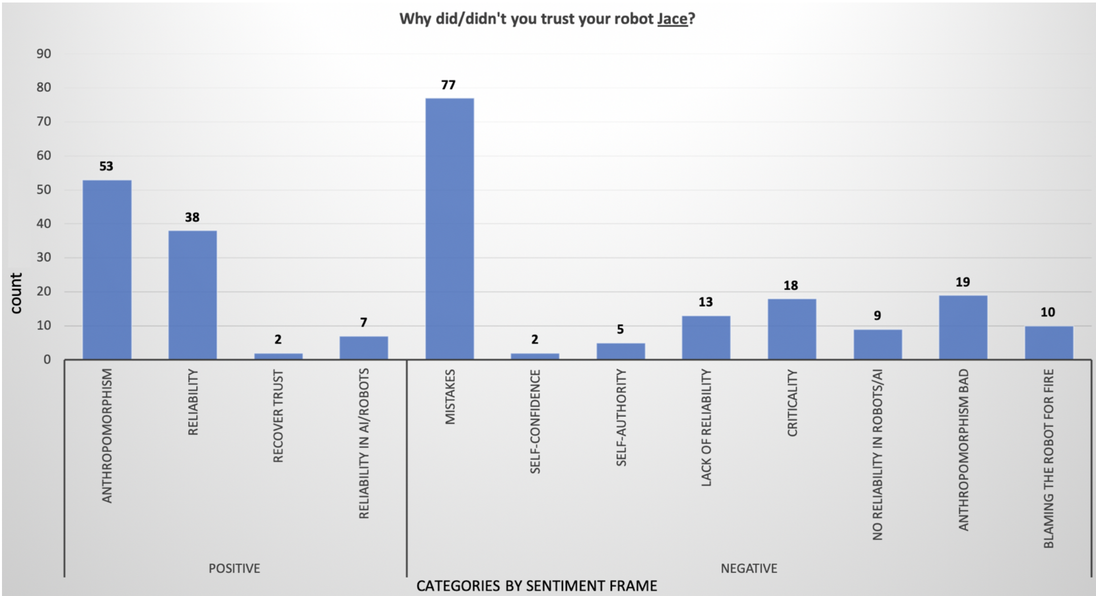


Figure 5: Qualitative analysis of participants' responses for the reasons why they did or did not trust the robot Jace. Categories are divided by differences in trusting response, positive and negative.

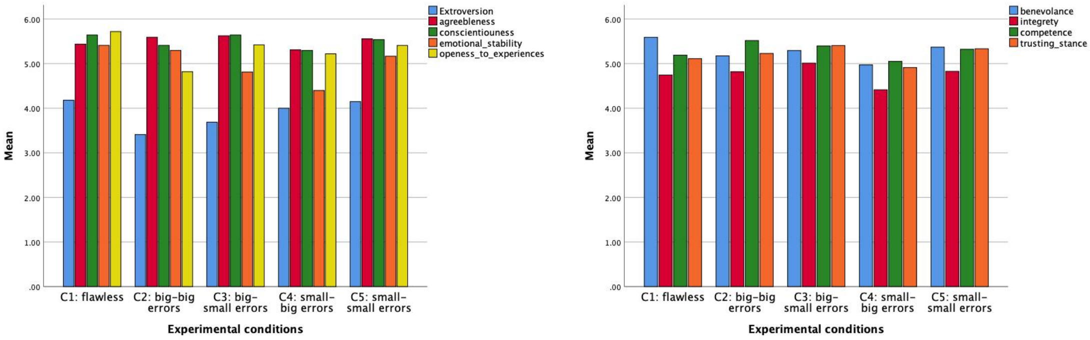
5.1 Effects of people's personality on trust

The participants' personal characteristics (i.e. personality traits and disposition of trust) for each experimental condition are shown in Figure 6. We can observe that the participants in our study, across the different experimental conditions, have similar personality traits and dispositions of trust. We did not find any statistical correlation between the experimental conditions and people's personality traits and disposition of trust. This means that any observed effects on participants' trust in different experimental conditions were not influenced by the distribution of participants.

A Cross-tabulation between the participants' disposition of trust and participants' personality traits shows that people's personality traits of agreeableness, conscientiousness and emotional stability are strongly connected to their disposition to trust other people ($p < 0.0001$).

Results of one-way ANOVA tests on participants' personality traits and their propensity of trusting the robot shown that participants' propensity for trusting the robot was correlated with conscientiousness trait ($p(3) = 0.042$, $F = 2.803$) and agreeableness trait ($p(3) = 0.022$, $F = 3.320$). We also observed that participants' benevolence trait was positively correlated with a higher trust in Jace ($p = 0.014$, $F = 6.078$).

This is in line with what is known in the literature for people with high agreeableness, conscientiousness and benevolence. According to Roccas et al. (2002), agreeableness exhibits



(a) *Participants' personality traits.* **(b)** *Participants' disposition of trust.*

Figure 6: Plots of participants' personal characteristics with respect to each experimental condition. a) Participants' personality traits, b)Participants' disposition of trust.

higher correlations with conformity, tradition and benevolence, and benevolence values correlated with trust, straight-forwardness, altruism and tender-mindedness facets. At the same time, agreeableness and conscientiousness correlate with life, work satisfaction and happiness, and people who tend to believe others are honest and trustworthy are more likely to trust others DeNeve and Cooper (1998).

5.2 Effects of people's past experiences on trust

We used a 7-point Likert Scales from 1 = "not at all" to 7 = "very much" to measure participants' experience with robots. The majority of the participants (75.97%) did not have any experience with robots when they joined the study (min = 1, max = 6, mean 1.64, std. dev. 1.27). Participants' past experiences with robots can be classified into the following four categories: 1) taking part in other user studies = 14.93%, 2) being a developer = 5.19%, 3) observing a robot = 11.68% and 4) being a researcher = 3.89%. They had experience with the following robots (multiple choice): industrial robots (e.g. robotic arms), virtual assistants, online/virtual interaction with robot, cleaning robots (e.g. Roomba), and watching robots in the media.

Analysing participants' past experiences with robots and their choices for trusting/not trusting Jace in the endangering scenario did not show any statistically significant correlation.

5.3 Effects of perception of robots

We categorised participants' responses to the Likert questions as negative when their ratings were less than 4, as moderate when the values were equal to 4, and as positive responses when their rating were greater than 4. Regarding the question "Would you feel comfortable having a robot as a companion in your home?", the majority of participants (69.48%) stated to be comfortable in having a robot as a robotic companion, 14.93% indicated that they neither agree nor disagree with the statement, while 15.58% did not want a robot in their homes.

The majority of participants (80.52%) expected to receive help from robots. Only 10.38% neither agreed nor disagreed with the statement, and 9.09% disagreed that they would expect help from a robot. We also noticed that participants who were more comfortable having a robot companion also expected to receive help from it (61.68%).

Participants also chose suitable roles for robots. Results indicate that 1) friend = 10.8%, 2) butler = 7.0%, 3) assistant = 24.6%, 4) tool = 18.6%, 5) companion = 11%, 6) pet = 6%, 7) machine = 13%. A few participants also wrote in the "other" option (0.2%) that robots should have a security role. We can observe that the majority of participants assigned the role of an assistant to a robot which is coherent with their expectations of receiving help from it. This is also in line with previous studies investigating the perceived role of a robot K. et al. (2005)

5.3.1 Perception of a robot as a companion

A Pearson correlation was run to determine the relationship between the participants' perception of a generic robot as a companion with both their experience with robots and participants' personality traits. We did not find any positive correlation, respectively with $p > 0.3, r = 0.082$, and $p > 0.04$, $r = 0.161$. On the contrary, participants with higher disposition of trust in peoples' benevolence were more comfortable with a robotic companion ($p = 0.039$, $r = 0.166$).

5.3.2 Expectation of a robot's capabilities

Participants' perception of usefulness of a generic robot was not correlated with their experience with robots ($p > 0.7, r = 0.026$). However, participants with a higher trusting stance ($p = 0.005$, $r = 0.227$) and belief of trusting people's competencies ($p = 0.011$, $r = 0.204$) expected robots to be helpful.

5.3.3 Perception of a robot's role

Mann-Whitney U-tests were performed to test the impact of the participants' prior experience and perceived role for robots. In particular, they were run to determine whether there were differences in participants' prior experiences score between those who selected or did not select a specific role. Results suggest that participants with a lower level of experience with robots tend to perceive them more as a machine ($p = 0.02, U = 1911$). Extroverted participants also perceived robots as machines ($p = 0.007$). In contrast, participants with a higher level of conscientiousness ($p = 0.040$) and agreeableness ($p = 0.007$) associated robots with a pet and an assistant. There was no statistically significant correlation between people's disposition of trust and the attributed robot's role.

5.4 Effects of perception of the robot Jace in relation to the magnitude of consequences of the errors

At the end of the interaction session, participants answered the same questions reported in the previous Section 5.3 related to the robot Jace that was used in this study.

5.4.1 Perceived companionship

In particular, we asked participants whether they wanted Jace or another robot as their home companion.

Spearman's rank-order analysis showed a positive correlation between participants desire of having Jace as home companion and both their the level of extroversion ($p = 0.001$, $r = 0.269$), and the level of trust in peoples' competencies ($p = 0.030$, $r = -0.175$). We also found a weak positive correlation that was statistically significant ($p = 0.05$, $F(154) = 0.156$) between the participants' level of experience with robots and their willingness of wanting the robot as home companion.

Further analysis found a statistically significant interaction between the effects of the level of participants' past experience with robots and their willingness of having the robot as companion across the five experimental conditions ($p(24) = 0.01, F = 1.952$). Observing the analysis, we identified a statistically significant difference in means between participants' previous experience with robots and their desire of having Jace as companion ($p < 0.0005$). However, simple main effects of participants' experience with robots on their acceptance of the robot as a companion showed a statistically significant difference when participants were tested in the flawless condition ($p(6, 32) = 0.005, F = 3.874$) and in the conditions with big errors ($p(6, 15) = 0.027, F = 3.326$). Analysing the participants' personalities and their desire of having Jace as home companion across the five experimental conditions, we observed a statistically significant correlation for participants who had higher level of agreeableness ($p(24) = 0.017$, $F = 1.839$), and emotional stability ($p(24) = 0.029$, $F = 1.727$).

A Spearman's rank-order analysis found also that participants' experience of robots affected their wish of having a robot different from Jace as a companion across the five experimental conditions ($p(22, 121) = 0.006, F = 2.084$).

5.4.2 Perceived reliability and faith in the ability of the robot

We foundn a correlation ($p(154) = 0.021, F = 0.186$) between the robot's perceived reliability and participants' experience of robots. Similarly, we find a statistically significant correlation ($p(154) = 0.004, F = 0.229$) between the participants' propensity to rely on the robot in uncertain and unusual situations and their previous experience with robots.

We also found a statistically significant interaction effect between people's familiarity with robots and their perceived reliability of the robot ($p = 0.04, F(50, 81) = 1.546$), and their propensity to rely on the robot in uncertain and unusual situations ($p = 0.001, F(51, 75) = 2.147$) according to the experimental conditions. In particular, we observed statistically significant differences between participants' experience with robots and their perceived reliability and participants' propensity of relying on the robot when participants were tested in the big-small error condition (respectively $p(32) = 0.018, F = 0.415$ and $p(32) = 0.046, F = 0.355$). These results are supported by de Graaf et al. (de Graaf and Ben Allouch, 2014) showing that a positive interaction with a robot can positively affect people's attitude towards robots. On the contrary, a negative experience with a robot can damage future interactions with other robots, as it appears to have happened in this study.

We observed a statistically significant correlation between the perceived reliability of the robot and people's level of extroversion trait ($p = 0.002$, $F = 2.729$), and between extroversion ($p(12) = 0.014$, $F = 2.214$) and emotion stability ($p(12) = 0.026$, $F = 2.025$) and people's reliability in the robot in uncertain and unusual situations.

5.4.3 Perception of the robot's role

A multiple linear regression analysis was run to predict participants' previous experience from their perceptions of the robot and the different experimental conditions. The condition where the robot showed flawless behaviours (condition **C1**) was used as the reference group for the multiple linear regression analysis.

We observed that participants with lower experience with robots perceived the robot as a friend ($p = 0.008$) and friendly ($p = 0.026$), but also as a toy ($p = 0.032$), when tested with the experimental condition with only small-errors.

We observed that participants perceived the robot as a friend ($p = 0.019$), and warm and attentive ($p = 0.025$) if they had a high level of extroversion, while those with lower extroversion perceived it as a machine ($p = 0.002$), when tested with the condition having severe errors at the beginning and small errors at end of the interaction (condition C3). Participants with high level of conscientiousness perceived Jace less as a friend in the condition in which the robot did big errors at the beginning and small errors at the end of the interaction (condition C3, $p = 0.0483$), but more as a butler in the other conditions ($p = 0.030$, $p = 0.001$, $p = 0.007$).

6 Study 2: Evolution of Trust and Erroneous Robot Behaviours

In this study, we were interested in investigating if the trust of humans in a robot can be recovered more easily if an error with severe repercussions happened at the beginning or end of repeated interactions. Here, we discuss the human-robot interactions observed in the study introduced in Section 3.2.

Results from theh previous study (described in Section 3.1) showed that participants' trust was affected more severely by the robot's errors with severe consequences, suggesting that participants tended to form their mental models of the robot at the beginning of interaction. However, study 2 is based on the assumption that people's trust can be recovered more easily when they already have an established bond with the robot, i.e. when the trust is broken at a later stage of the interaction (Schilke et al., 2013; Bottom et al., 2002).

Participants unanimously judged the level of realism of the scenarios with ratings higher than five on a 7-point Likert Scale, ranged from 1 to 7 (disagree to agree).

6.1 Trust in Care-O-bot 4 in relation to the robot's errors

Participants trusted the robot more when they were experiencing robot's behaviours with big errors at the end of the interaction (condition **CP1**), compared to when the errors were made at the beginning of the interaction. Two out of three participants trusted the robot to be able to handle the emergency situation, and one preferred to deal with the emergency situation in collaboration with the robot when tested in condition **CP1**. When tested in condition **CP2**, participants often did not trust the robot (1 out of 3 participants), and did not trust either themselves or the robot (2 out of 3 participants).

A participant tested with condition **CP2** rushed towards the house's entrance (i.e. to exit the house) being scared of the emergency situation, while a participant in condition **CP1** blamed the robot for the fire. Another participant asked the robot for a fire extinguisher, and invited the robot to call the fire brigade.

We also asked participants to justify their choices of trusting or not trusting the robot using an open-ended question. Their answers highlighted that their trust for the robot was based on the idea that the robot earned it during the interaction, or according to their propensity of trusting others. For example, participants stated that “I easily trust everyone” and “I believe that people know what they are doing”. A participant also commented that “he (*the robot*) was correct all the time”. In contrast, participants did not trust the robot due to its limited capabilities, i.e. in movements, dialogues, etc. Some commented that “I would trust him (*the robot*) with most things” and “I personally trust in robot with regular tasks such as reminding, cleaning”, and “he (*the robot*) would understand my commands correctly”, or “the responding of the Care-O-bot still slow and not precise”. Moreover, one participant was particularly concerned by the robot revealing their secret, and commented “it (the robot) promised not to tell my secret”. Figure 7 summarises the qualitative analysis ran on participants’ answers to the open-ended question.

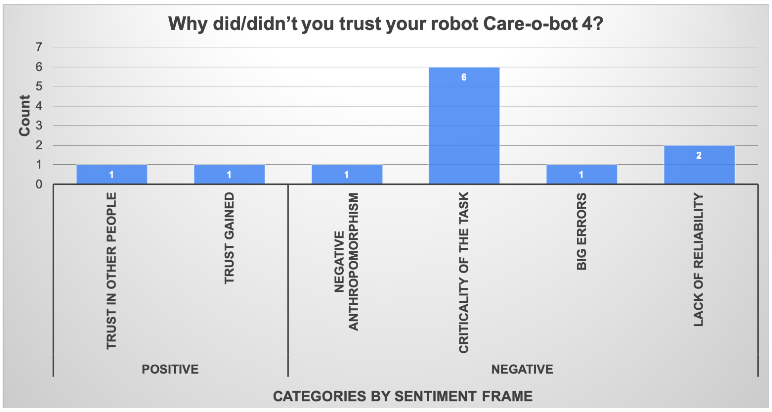


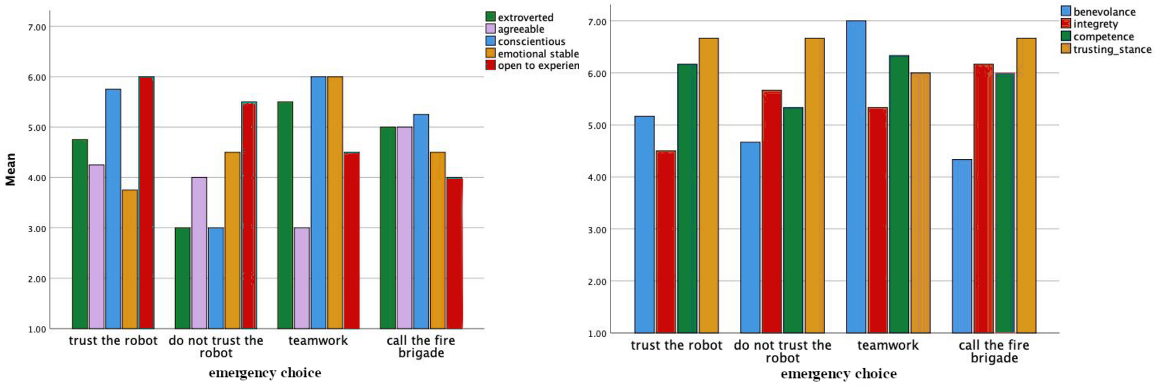
Figure 7: The participants’ motivations for trusting or not trusting the robot are here summarised according to positive and negative categories.

6.2 Antecedents of trust

We studied the effects of participants’ antecedents of trust (past experiences, personality traits and disposition of trust) on their choices of trust in the robot.

Participants did not have any, or very limited (i.e. participants in other studies), previous experience with robots.

As shown in Figure 8, there was no difference between participant’s choice of trust (in the emergency scenario) and the distribution of their personalities, and the distribution of their disposition to trust others. However, we can observe that the participants with higher conscientiousness (Figure 8a), openness to experience, competence and trusting stance (Figure 8b) also trusted the robot. The participant with lower extroversion, conscientiousness, benevolence, and higher trusting stance trusted to work with the robot. Participants with high conscientiousness, emotional stability, benevolence and competence did not trust the robot.



(a) *Distribution of personality traits.* **(b)** *Distribution of disposition of trust.*

Figure 8: Distribution of participants' (a) personality traits and (b) disposition of trust by trust choice for each of their emergency choice.

We acknowledge that the very small number of participants cannot give us a high degree of confidence, and we will need to consider to further investigate these effects with a larger and more diverse group. However, one-way ANOVA tests showed that there was no statistical significant difference between participants' choice to trust the robot and their personal traits and disposition: extroversion ($p > 0.05$), agreeableness ($p > 0.05$), conscientiousness ($p = 0.05$), emotional stability ($p = 0.05$) and open to experience ($p > 0.05$), benevolence ($p > 0.05$), integrity ($p > 0.05$), competence ($p = 0.5$) and trusting stance ($p > 0.05$).

6.3 Perception of Care-O-bot 4

As we did in Study 1, we asked participants whether they would have wanted the robot as their home companion, and which were the roles considered suitable for the robot.

The majority of participants (4 out of 6 participants) stated to want Care-O-bot 4 as their robotic companion, while the remaining two participants were not positive or unsure to have the robot in their homes.

Participants had varied opinions on the suitable role for the robot. Two of the 6 participants perceived the robot as an assistant, the remaining four participants perceived the robot either as a tool, a companion, a friend, or a butler.

Measuring people's perception of reliability and faith to perform correctly in unexpected situations, we observed that participants trusted Care-O-bot 4 ($n = 4.17 \pm 0.24$) less in condition **CP2** than condition **CP1**, while participants in condition **CP1** decided to call the fire brigade ($n = 6.17 \pm 0.24$).

6.4 Evaluation of robot errors

At the end of the conditions, the robot verbally asked the participants to state whether it made any errors, and to select the scenarios that contain robot's errors. Participants provided responses to both questions by selecting them on the robot's tablet.

In their responses, participants stated that the robot did not make any mistakes. However, when we asked them to rate the errors (including the ones made by the robot during the interaction), according to their level of consequences (i.e. with severe consequences), the resulting rankings confirmed our expectations. We believe that they stated that the robot did not make any errors due to a possible bystander effect, which might have inhibited participants to express an open negative consideration in the presence of the culprit Voelpel et al. (2008) (i.e. the robot in our study).

Finally, we asked participants to identify the scenarios they would entrust the robot to deal with, in scenarios different from the fire emergency. They stated to not trust the robot to be able to take care of life-threatening scenarios, such as "If your beloved ones were in life-danger, would you trust me to deal with it?", but they trusted the robot to be able to handle cognitive and lower risks situations, such as "If you needed to take medicines regularly, would you trust me to remind you of taking them?", or to remind them of important meetings, and to manage a smart house such as Robot House.

7 Limitations

The results of the two studies presented in this article highlight the various factors that can affect people's trust in robots. However, there are several limitations.

Over the last decade, online surveys, questionnaires and experiments have become standard tools to conduct research both in Academia Sheehan and Pittman (2016) and Industry thanks to the use of web services, as SurveyMonkey and Amazon Mechanical Turk that increase the efficiency and effectiveness of the data gathering process Buhrmester et al. (2011).

Studies conducted through crowdsourcing services can collect participants' responses very fast. This might imply that the percentage of diversity of participants might change depending on the time zone of the users of the crowdsourcing services. Future research should consider investigating whether collecting responses of smaller groups of participants, by publishing the recruitment according to different time zones would yield a wider diversity of the sample.

While the interactive scenarios were perceived by participants as very realistic and immersive, participants might well have a different mindset in a real situation when meeting a robot 'face to face', and where a prompt reaction may be needed or expected. Moreover, investigating people's trust in robots in real life-threatening scenario can be a challenging task due to ethical and legal concerns (Salem and Dautenhahn, 2015).

Finding participants for in-person studies is extremely difficult. In particular, when investigating long-term effects and changes over time in HRIs whn participants are asked to attend many sessions over several weeks. We were able to consistently establish that robot behaviours affected participants' trust in them. However, larger scale trials need to consolidate these findings, and also provide further insights to unravel the complexity of trust dynamics between humans and robots.

8 Conclusion & Future Work

The following research questions emerged from our review of related work when investigating the trust dynamics between humans and robots:

RQ-1 How do various types of robot errors affect human's trust in a robot?

RQ-2 How does people's trust in a robot change according to their personal differences?

RQ-3 Does people's trust in a robot change over time if the initial conditions (positive or negative) of trust in the robot changes?

In this work, we presented two studies used to answer these questions.

RQ-1 How do various types of robot errors affect human's trust in a robot? We used an interactive storyboard presenting ten different scenarios in which a robot completed tasks under five different conditions to explore the first two research questions. Results showed that participants' trust was affected more severely when the robot made errors with severe consequences.

RQ-2 How does people's trust in a robot change according to their personal differences? While analysing people's individual differences, we found that participants' individual traits are correlated with their perception and trust of the robot. A strong relationship was found between participants' personalities (agreeableness, conscientiousness and emotional stability) and their disposition to trust other people. The robot was perceived as a friend, warm and attentive by extroverted participants, while it was considered a tool by more participants. We also found that the extroversion trait affected participants' desire of having the robot as home companion, and their beliefs in its reliability and trustworthiness in uncertain and unusual situations. We found that conscientiousness and agreeableness traits correlate with participants' propensity for trusting the robot. Participants' belief in benevolence of people also correlate with higher trust in the robot.

The majority of the participants did not have experience with robots. We observed that people who had negative previous experience with a robot were less inclined to trust the robot that made big errors in our studies, while a positive experience with a robot consequently affected people's positive predisposition towards a robot that behaved flawlessly.

RQ-3 Does people's trust in a robot change over time if the initial conditions (positive or negative) of trust in the robot changes? Study 2 investigated if people would trust a robot that broke their trust in an initial or later stage of the interaction. The findings showed that people's trust was affected the most when the robot made errors at the beginning of the interaction. Moreover, people's lack of trust in the robot was also connected to the criticality of the task undertaken by the robot. These results corroborate the known belief that people's reliability in a robot is also affected by the possibility of a negative outcome (Mayer et al., 1995; Lee and Moray, 1992).

8.1 Original contributions to knowledge

Robot errors have been shown to reduce the perceived reliability and trustworthiness of robots in several studies Desai et al. (2013); Hancock et al. (2011b); Salem et al. (2015). These works highlighted that users complied with the robots' directions and suggestions discarding previous robotic failures Robinette et al. (2016); Salem et al. (2015); Bainbridge et al. (2011). These studies also were characterised by the fact that the robots' errors were not distinguished by a different magnitude of consequences. In this paper, we have shown that errors with severe consequences affected people's trust in robots more than errors with minor consequences.

Corritore et al. (2003) have shown that a sequence of small errors can affect people's trust in robots more severely and for a longer period than one single big error. In section 3.2, we have shown that the timing in which the errors occurs may impact people's trust in robots differently, particularly in the case of the robot making errors that have major consequences. Our findings also suggest that participants' judgements on whether to trust or not to trust the robot are principally formed after a few initial interaction sessions with the robot, which is inline with the finding from Yu et al. (2017).

Moreover, we have shown that people's perception of the robot and its errors consequently also affect their trust. Indeed, the findings showed that individuals' personality traits and personal dispositions, and previous experiences with robots influenced their trust in the robot, particularly, when the robot was making big errors.

8.2 Future works

The insights gained by this research have shown that it is possible to build a successful collaboration between people and robots based on trust. However, they have also opened up new directions for investigating trust in HRI, and identified a number of future challenges to overcome.

In our investigations, we outlined several similarities and differences between the virtual and real (in-person) studies. However, the unbalanced sample sizes do not allow us to make a more extensive comparison between the two sets of results. In future, it would be useful to address the samples sizes, to further investigate the possibility of using virtual setup to help to assess in-person HRI, and to identify commonalities, as well as phenomena that would only emerge uniquely in virtual or in person HRI.

The research presented in this article highlighted the necessity of further understanding how human-robot relationships are formed, and which robot factors, including familiarity, appearance and perception as social entity, will influence most people's trust in robots. Indeed, in study 2 (see Section 6) we observed that participants were reluctant to communicate their disapproval of the robot for its errors. This most probably happened due to the effect well-known in psychology and human-computer interaction (i.e., bystander effect or social inhibition of helping). It seems to have milder effects in online interactions Chekroun and Brauer (2002). Future research should investigate to what extent people's mental models, including the perceived implications of task outcomes and consequences on their persona, inhibits their behaviours in the presence of robots.

The results of this research have found that people's previous experiences of robots, personality traits, and dispositions to trust humans affects their trust in robots. However, people are now becoming surrounded by digital technologies, and it is difficult to match people's expectations of robots with their experience with more robust and advanced AIs, such as Alexa or Google Assistant. Further studies should aim to integrate modern learning techniques (i.e. convolutional neural networks, deep learning etc.) that allow more fluid and rich interactions in HRI studies in order to further investigate how people's perceptions of robots affects their trust in it. This will contribute to develop robots that adapt to interact naturally with people.

References

Adams, B. D., Bruyn, L. E., Houde, S., Angelopoulos, P., Iwasa-Madge, K., and McCann, C. (2003). Trust in automated systems. *Ministry of National Defence*.

Agresti, A. (2002). *Categorical data analysis*. Wiley-Interscience, Chichester;New York;, 2nd edition.

Ambady, N., Bernieri, F. J., and Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32:201–271.

Aroyo, A. M., Pasquali, D., Kothig, A., Rea, F., Sandini, G., and Sciutti, A. (2021). Expectations vs. reality: Unreliability and transparency in a treasure hunt game with icub. *IEEE Robotics and Automation Letters*, 6(3):5681–5688.

Atkinson, D. J., Clancey, W. J., and Clark, M. H. (2014). Shared awareness, autonomy and trust in human-robot teamwork. In *In Artificial Intelligence and Human-Computer Interaction: Papers from the 2014 AAAI Spring Symposium on.*

Bainbridge, W. A., Hart, J. W., Kim, E. S., and Scassellati, B. (2011). The benefits of interactions with physically present robots over video–displayed agents. *International Journal of Social Robotics*, 3(1):41–52.

Barrick, M. R., Stewart, G. L., Neubert, M. J., and Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, 83(3):377–391.

Bottom, W. P., Gibson, K., Daniels, S. E., and Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, 13(5):497–513.

Buhrmester, M., Kwang, T., and Gosling, S. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6:3–5.

Chekroun, P. and Brauer, M. (2002). The bystander effect and social control behavior: the effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology*, 32(6):853–867.

Colquitt, J. A., Scott, B. A., and LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4):909–927.

Corritore, C. L., Kracher, B., and Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International Journal of Human - Computer Studies*, 58(6):737–758.

Costa, A. C., Roe, R. A., and Taillieu, T. C. B. (2001). Trust implications for performance and effectiveness. *European Journal of Work and Organizational Psychology*, 10(3):225–244.

Daniela, C., E, C., and S, B. (2017). Personality factors and acceptability of socially assistive robotics in teachers with and without specialized training for children with disability. *Life Span and Disability*, 20(2):251–272.

de Graaf, M. M. and Ben Allouch, S. (2014). Expectation setting and personality attribution in hri. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, page 144–145, New York, NY, USA. Association for Computing Machinery.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., and Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, page 459–478.

DeNeve, K. M. and Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, 124:197–229.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 251–258.

Deutsch, M. (1958). Trust and suspicion. *The Journal of Conflict Resolution*, 2(4):265–279.

Drury, J. L., Scholtz, J., and Yanco, H. A. (2003). Awareness in human-robot interactions. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 1, pages 912–918.

Elson, J. S., Derrick, D. C., and Ligon, G. S. (2018). Examining trust and reliance in collaborations between humans and automated agents. In *HICSS*.

Freedy, A., DeVisser, E., Weltman, G., and Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS*, pages 106–114.

Furlough, C., Stokes, T., and Gillan, D. J. (2019). Attributing blame to robots: I. the influence of robot autonomy. *Human Factors*, 0(0):1–11.

Gockley, R. and Matariundefined, M. J. (2006). Encouraging physical therapy compliance with a hands-off mobile robot. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, page 150–155, New York, NY, USA. Association for Computing Machinery.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B., J. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528.

Haberlandt, K. (1982). Reader expectations in text comprehension. In Ny], J.-F. L. and Kintsch, W., editors, *Language and Comprehension*, volume 9 of *Advances in Psychology*, pages 239–249. North-Holland.

Hancock, P. A., Billings, D. R., and Schaefer, K. E. (2011a). Can you trust your robot? *Ergonomics in Design*, 19(3):24–29.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011b). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 53(5):517–527.

Haring, K. S., Matsumoto, Y., and Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Lecture Notes in Engineering and Computer Science*, 1:425–430.

Haselhuhn, M. P., Schweitzer, M. E., and Wood, A. M. (2010). How implicit beliefs influence trust recovery. *Psychological Science*, 5:645–648.

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

HT, R., MR, M., PA, C., PW, E., and EJ, F. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of Personality and Social Psychology*, 101(3):557-570.

K., D., S., W., C., K., M.L., W., K.L., K., and I., W. (2005). What is a robot companion - friend, assistant or butler? In *Procs IEEE/RSJ Int Conf on Intelligent Robots and Systems 2005*, pages 1488–1493.

Koay, K., Walters, M., May, A., Dumitriu, A., Christianson, B., Burke, N., and Dautenhahn, K. (2013). Exploring robot etiquette: Refining a hri home companion scenario based on feedback from two artists who lived with robots in the uh robot house. In *Social Robotics*, Lecture Notes in Computer Science, pages 290–300. Springer. 5th Int Conf on Social Robotics, ICSR 2013 ; Conference date: 27-10-2013 Through 29-10-2013.

Kudryavtsev, A. and Pavlodsky, J. (2012). Description-based and experience-based decisions: Individual analysis. *Judgment and Decision Making*, 7(3):316–331.

Lee, A. Y. (2001). The mere exposure effect: An uncertainty reduction explanation revisited. *Personality and Social Psychology Bulletin*, 27(10):1255–1266.

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

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80.

Lee, M. K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., and Savetsila, S. (2012). Personalization in hri: A longitudinal field experiment. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 319–326, New York, NY, USA. Association for Computing Machinery.

Lemaignan, S., Fink, J., Mondada, F., and Dillenbourg, P. (2015). You're doing it wrong! studying unexpected behaviors in child-robot interaction. In Tapus, A., André, E., Martin, J.-C., Ferland, F., and Ammi, M., editors, *Social Robotics*, pages 390–400, Cham. Springer International Publishing.

Lewis, J. D. and Weigert, A. (1985). Trust as a social reality. *Social Forces*, 63(4):967–985.

MacDorman, K. F. and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337.

Macias, W. (2003). A beginning look at the effects of interactivity, product involvement and web experience on comprehension: Brand web sites as interactive advertising. *Journal of Current Issues & Research in Advertising*, 25(2):31–44.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, pages 709–734.

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

McKnight, D. H., Choudhury, V., and Kacmar, C. (2001). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19(2):98–100.

Muir, B. M. and Moray, N. (1996). Trust in automation: Part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39:429–460.

Paetzel, M., Perugia, G., and Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI 20, page 73–82, New York, NY, USA. Association for Computing Machinery.

Pixar (2011). The moon, short movie. https://www.youtube.com/watch?v=vbuq7w3ZDUQ. last accessed in May 6th, 2020.

Rae, I., Takayama, L., and Mutlu, B. (2013). In-body experiences: Embodiment, control, and trust in robot-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1921–1930, New York, NY, USA. Association for Computing Machinery.

Reber, R., Winkielman, P., and Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1):45–48.

Robert, L. P. (2018). Personality in the human robot interaction literature: A review and brief critique. In *Proceedings of the 24th Americas Conference on Information Systems*.

Robinette, P., Howard, A. M., and Wagner, A. R. (2015). Timing is key for robot trust repair. *Social Robotics. Lecture Notes in Computer Science*, 9388:574–583.

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *Proceeding HRI '16 The Eleventh ACM/IEEE International Conference on Human Robot Interation*, pages 101–108. IEEE Press Piscataway.

Roccas, S., Sagiv, L., Schwartz, S. H., and Knafo, A. (2002). The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

Ross, J. M. (2008). Moderators of trust and reliance across multiple decision aids (doctoral dissertation), university of central florida, orlando.

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2017a). How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In Kheddar, A., Yoshida, E., Ge, S. S., Suzuki, K., Cabibihan, J.-J., Eyssel, F., and He, H., editors, *Social Robotics*, pages 42–52, Cham. Springer International Publishing.

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2017b). Human perceptions of the severity of domestic robot errors. In Kheddar, A., Yoshida, E., Ge, S. S., Suzuki, K., Cabibihan, J.-J., Eyssel, F., and He, H., editors, *Social Robotics*, pages 647–656, Cham. Springer International Publishing.

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2018a). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn Journal of Behavioral Robotics*, 9.

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2020a). How social robots influence people's trust in critical situations. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1020–1025.

Rossi, A., Holthaus, P., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2018b). Getting to know pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI '18, pages 246–252, New York, NY, USA. ACM.

Rossi, A., Moros, S., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2019). Getting to know kaspar : Effects of people's awareness of a robot's capabilities on their trust in the robot. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–6.

Rossi, A. and Rossi, S. (2021). Engaged by a bartender robot: Recommendation and personalisation in human-robot interaction. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 115–119, New York, NY, USA. Association for Computing Machinery.

Rossi, S., Rossi, A., and Dautenhahn, K. (2020b). The secret life of robots: Perspectives and challenges for robot's behaviours during non-interactive tasks. *International Journal of Social Robotics*.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust1. *Journal of Personality*, 35(4):651–665.

Sadeghi, N., Kasim, Z. M., Tan, B. H., and Abdullah, F. S. (2012). Learning styles, personality types and reading comprehension performance. *Psychology*.

Salem, M. and Dautenhahn, K. (2015). Evaluating trust and safety in hri: Practical issues and ethical challenges.

Salem, M., Lakatos, G., Amirabdollahian, F., and 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*, HRI 15, page 141–148, New York, NY, USA. Association for Computing Machinery.

Schilke, O., Reimann, M., and Cook, K. S. (2013). Effect of relationship experience on trust recovery following a breach. *Proceedings of the National Academy of Sciences*, 110(38):15236–15241.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., and Young, J. E. (2015). Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 125–132, New York, NY, USA. Association for Computing Machinery.

Sheehan, K. and Pittman, M. (2016). The academic's guide to using amazon's mechanical turk: The hit handbook for social science research. *Irving: Melvin & Leigh*.

Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226.

Takayama, L. and Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS 09, page 5495–5502. IEEE Press.

Tannenbaum, K. R., Torgesen, J. K., and Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, 10(4):381–398.

Tseng, S. H., Hua, J. H., Ma, S. P., and e. Fu, L. (2013). Human awareness based robot performance learning in a social environment. In *2013 IEEE International Conference on Robotics and Automation*, pages 4291–4296.

van Maris, A., Lehmann, H., Natale, L., and Grzyb, B. (2017). The influence of a robot's embodiment on trust: A longitudinal study. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, page 313–314, New York, NY, USA. Association for Computing Machinery.

Voelpel, S. C., Eckhoff, R. A., and Förster, J. (2008). David against goliath? group size and bystander effects in virtual knowledge sharing. *Human Relations*, 61(2):271–295.

Walters, M. L., Oskoei, M. A., Syrdal, D. S., and Dautenhahn, K. (2011). A long-term human-robot proxemic study. In *IEEE RO-MAN*, pages 137–142.

Williamson, J. M. (2018). Chapter 1 - individual differences. In Williamson, J. M., editor, *Teaching to Individual Differences in Science and Engineering Librarianship*, pages 1–10. Chandos Publishing.

Wood, T. (2014). Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Educ Theory Pract*, 19(3):409-427.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. *ACM*, 126745:307–317.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2):1–27.