Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust

Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, Kerstin Dautenhahn
University of Hertfordshire
College Lane, Hatfield
AL10 9AB, United Kingdom
{m.salem, g.lakatos, f.amirabdollahian2, k.dautenhahn}@herts.ac.uk

ABSTRACT

How do mistakes made by a robot affect its trustworthiness and acceptance in human-robot collaboration? We investigate how the perception of erroneous robot behavior may influence human interaction choices and the willingness to cooperate with the robot by following a number of its unusual requests. For this purpose, we conducted an experiment in which participants interacted with a home companion robot in one of two experimental conditions: (1) the correct mode or (2) the faulty mode. Our findings reveal that, while significantly affecting subjective perceptions of the robot and assessments of its reliability and trustworthiness, the robot's performance does not seem to substantially influence participants' decisions to (not) comply with its requests. However, our results further suggest that the nature of the task requested by the robot, e.g. whether its effects are revocable as opposed to irrevocable, has a significant impact on participants' willingness to follow its instructions.

Keywords

Social Human-Robot Interaction; Cooperation; Trust

1. INTRODUCTION

Robots are increasingly being developed for use in social settings, e.g. to assist humans at work or at home, both with everyday tasks and in healthcare scenarios. For example, a home companion robot could remind an elderly person to take their medication or to engage in regular physical exercise. In the domestic domain, such interactions are typically expected to take place in an informal and unstructured way, resulting in numerous challenges when designing robots intended to interact socially in these complex environments. In addition to work on technical reliability, this has motivated different lines of research into the factors that may impact the quality of social human-robot interaction (HRI) and acceptance of the robotic assistant itself. One factor of crucial importance when establishing and maintaining effective relationships with robots is *trust* [10]. Playing a major role in human interactions, especially with regard to critical decisions, trust is similarly believed to increase a robot's capacity to be accepted as a collaborative partner [14]. Trust is fundamental in social contexts, as it is tightly linked to persuasiveness and can directly affect people's willingness to accept information provided by the robot and to follow its suggestions [8]. Thus, it is desirable to design robots that act socially in a way such that humans can develop trust toward them and cooperate with them. In view of robotic helpers assisting humans in their homes in the not-too-distant future, an important research question is how to make robots trustworthy to assist non-expert users and thereby increase their acceptance, persuasiveness and likability.

The present work aims to explore factors that may affect how humans perceive and the extent to which they are willing to 'trust' a robotic assistant based on its exhibited cognitive and behavioral skills. Our experimental design partly draws inspiration from a study presented by Bainbridge et al. [1], which measured whether human participants trusted a robot by following its 'unusual request' of throwing away a pile of new textbooks in someone's office, either based on the robot's physical versus on-screen presence. However, in our work the focus does not lie on effects of the robot's level of embodiment, but on the role that errors made by the robot might play when establishing human-robot trust.

2. BACKGROUND AND MOTIVATION

To date, trust is still a fairly underrepresented line of HRI research, which is partly due to the complexity of the concept itself: although trust has been studied in a wide range of disciplines (e.g. psychology, sociology, philosophy, economics), each discipline relies on its own definitions and findings which often lack agreement and generalization [4].

2.1 Trust in Human-Machine Interaction

Of greater relevance to trust in HRI and already more extensively studied, previous research on trust in automation, e.g. [17, 18], and in human-computer interaction (HCI), e.g. [16, 3], may provide some insights and implications for trust in HRI. However, robots differ from automated machines and computer interfaces in that they are mobile and of a greater degree of embodiment, e.g. in order to fulfill their designated social and operative functions. As a result, interaction with a robot is potentially richer: humans can, for example, walk around or touch a real robot, which in turn results in a different dimension of risks and safety concerns. These dissimilarities could suggest that human trust may vary for robots compared to automation or even HCI.

HRI'15, March 2–5, 2015, Portland, Oregon, USA.
ACM 978-1-4503-2883-8/15/03.
http://dx.doi.org/10.1145/2696454.2696497.

Although hardly any direct comparisons have been made between trust in automation, HCI and HRI, findings from the first two domains can serve as starting points to identify factors that may influence humans' trust in robotic agents. For example, in all three areas both underreliance and overreliance caused by inappropriate levels of trust can result in dissatisfying human-machine interaction [10].

A consistent definition of trust has not emerged in the automation or HCI literature, however, most concepts of trust are multidimensional and include reliability and predictability as some of the promoting factors. Muir and Moray [17] argue that trust is based mostly on the extent to which the machine is perceived to perform its function properly, suggesting that machine errors strongly affect trust. Although the magnitude of an error is an important factor regarding the loss of trust, an accumulation of small errors seem to have a more severe and long-lasting impact on trust than a single large error [4]. In contrast, however, previous work in HRI (e.g. [19]) has found that errors occasionally performed by a humanoid robot can actually increase its perceived humanlikeness and likability.

Consequently, one pushing question is whether findings from automation and HCI can be transferred to HRI, that is, how erratic behavior can affect a social robot's perceived trustworthiness as well as people's willingness to cooperate with it. Therefore, our present work sets out to shed light on the process of trust development in social HRI.

2.2 Measuring Trust in HRI

Measuring trust in HRI is not a straightforward task. Hancock et al. [10] find most reviews of human-robot trust to be rather qualitative and descriptive, mainly measuring a momentary state of trust instead of the process of trust development and the factors involved in it. In their quantitative review of the existing body, they reveal that robot characteristics, in particular with regard to its performance, are the most influential drivers of perceived trust in HRI. But only few if any HRI studies have systematically investigated the role of human-related characteristics (e.g. level of expertise, personality traits such as extroversion [11]) and environmental factors (e.g. culture, task type [15]).

A substantial portion of related work (e.g. [14, 11]) employ so-called economic trust games to measure the level of trust placed in an agent. However, since these games only model very specific trust situations related to monetary gain or loss, findings from such studies cannot be easily generalized. More importantly, in many studies, trust is measured solely with regard to one single task context, thus not allowing for a comparison in case the effects would deviate in a different task or situation. Therefore, one of the major challenges when investigating trust in social HRI is to design study scenarios that demand trust in a natural and realistic environment, while ideally incorporating a variety of tasks which tap different dimensions of trust.

Since there may be discrepancies between subjective (self-reported) and objective (behavioral) sources of data, a number of studies have combined both (e.g. [1]), and our present work follows the same principle. Drawing inspiration from [1], in our current study we measure trust based on self-reported quantitative and qualitative questionnaire data as well as on behavioral data that assesses trust as the participants' *willingness to cooperate* [14] with a robot when it addresses them with a number of usual and unusual requests. In this way, our behavioral measure is based on HCI-related research that defines cooperation as a "behavioral outcome of trust" [21].

3. METHOD

We conducted an experiment to gain a deeper understanding of how a robot's faulty behavior might impact and shape human experience and evaluation of HRI. For this, we investigated both subjectively self-reported and objectively measured behavioral effects based on different interaction tasks.

3.1 Hypotheses

Based on findings from related work on trust in Psychology, automation, HCI and HRI we developed three main hypotheses for our experiment:

1. *Effect of condition.* Manipulation of the robot's behavior (correct vs. faulty performance) will affect
   (a) participants' perception of the robot and the interaction (subjective assessment of HRI).
   (b) participants' performance when cooperating with the robot (objective assessment of HRI).
2. *Effect of type of task request.* The nature of the task will have an effect on participants' willingness to follow the robot's instructions.
3. *Effect of participant's personality.* Participants' personality traits (e.g. extroversion) will affect
   (a) participants' perception of the robot and the interaction (subjective assessment of HRI).
   (b) participants' willingness to collaborate with the robot (objective assessment of HRI).

3.2 Experimental Design

We conducted a between-participants experimental study in which participants interacted with the Sunflower Robot [13], a mobile non-humanoid robot consisting of a Pioneers platform with an embodied upper body (see **Figure 1b**). The robot can navigate autonomously while relying on a range of sensors to avoid collisions with humans and objects such as furniture. Rather than using a modified laboratory, the experiment took place in a realistic domestic environment, i.e. a regular three-bedroom house near the University of Hertfordshire, UK, which has been equipped with various sensing devices and is frequently used for human-robot interaction studies in the home care context.

We manipulated the robot's behavior in two experimental conditions: the **correct** *(C)* and the **faulty** *(F)* mode. In condition *C*, the robot correctly translated user input into action and navigated in a smooth and goal-directed manner. In condition *F*, the robot showed cognitive and physical imperfections, e.g. by incorrectly "remembering" a user selection and by navigating in an erratic manner, i.e. by moving into the wrong direction and occasionally spinning around itself. By comparing the effect of experimental condition we aimed to gain insights into the mental processes that drive a human user's decision to trust a robotic agent.

3.3 Experimental Procedure

Participants were tested individually. They were greeted by the experimenter at the house entrance and led into the living room area to receive a brief description of the experimental process. After reviewing and signing a consent form,

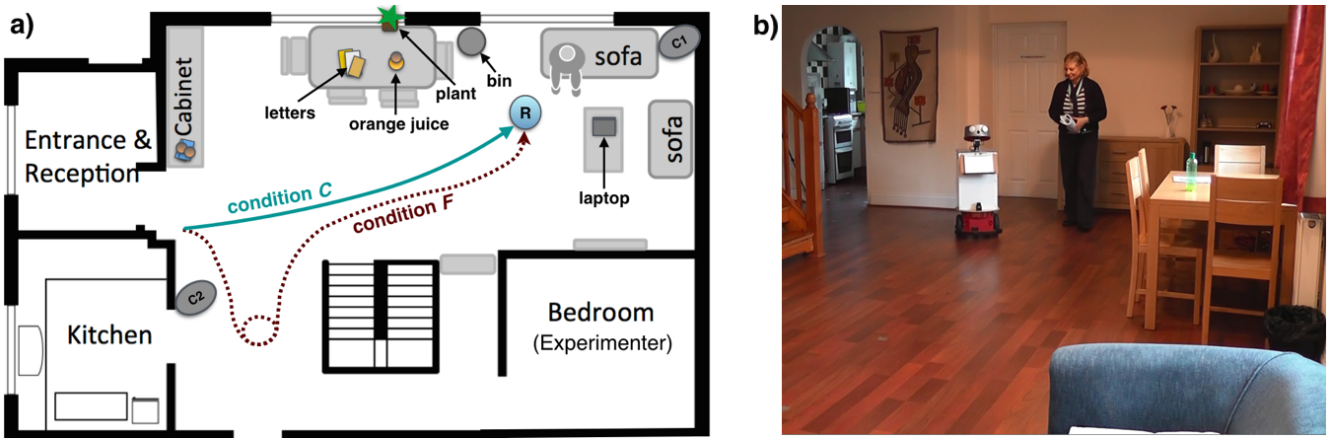


Figure 1: Experimental setting: a) schematic drawing including robot's sample navigation path in correct vs. faulty condition; b) snapshot perspective from camcorder C1

they were asked to fill out a questionnaire recording their demographic background, previous experience with robots and technology and tapping some personality traits as well as their attitudes and expectations regarding robots.

Participants were then introduced to the study scenario and the interaction. They were told that they are visiting a friend at home to prepare and have lunch together. The friend's robotic assistant would welcome them at the door. They were instructed to interact with the robot as naturally as possible and in a way that feels comfortable to them. All further required information would be provided to them in the course of the interaction. To communicate with the participant during the interaction, the robot displayed messages on a tablet attached to its torso, which were accompanied by flashing LED lights to attract the participant's attention.[1]

The experiment consisted of two interaction stages: *demonstration of competence* stage and *unusual requests* stage. The first stage aimed to demonstrate the robot's level of cognitive and physical competence. That is, in condition *C*, the robot showed its 'flawlessness' by avoiding mistakes and by exposing goal-directed and legible navigation, whereas in the *F* mode, its 'imperfections' were demonstrated by faulty behaviors and illegible navigation (i.e. occasionally navigating into the wrong direction first or spinning around). In the second stage, the robot asked the participant to perform actions that may appear unusual so they might hesitate to comply with the requests. This was to measure the participant's trust regarding the robot and the legitimacy of its requests, as well as their willingness to cooperate with it.

Demonstration of Competence Stage

Greeting: The robot greeted the participant at the entrance by displaying the message "Welcome to our house. Unfortunately, my owner has not returned home yet. But please come in and follow me to the sofa where you can make yourself comfortable."

Going to Sofa: Once the participant had confirmed reading this message by pressing OK, the robot guided the person to the sofa. In condition *C*, the robot navigated to the sofa on the direct route, whereas in condition *F*, it took off into the wrong direction, stopped abruptly before spinning around itself a few times and finally moving towards the sofa (see schematic path illustration in **Figure 1a**).

Offering to Play Music: Once the person had sat down on the sofa, the robot displayed the message "Would you like to listen to some music?" together with the options "Yes, Classical Music", "Yes, Rock Music" and "No, thanks". In *C*, the robot then started playing the correct type of music, or nothing if the option "No, thanks" had been selected. In *F*, the robot played the other - wrong - type of music, or classical music if "No, thanks" had been selected.

Set-up Table: The robot displayed the message "Would you be so kind as to help me setup the table?", and once the participant had pressed OK: "Please follow me to the cabinet." The robot then navigated to the cabinet on which a tablecloth, two plates, two cups and cutlery could be found, and presented its tray before displaying the message: "Please put the tablecloth, plates, cups and cutlery on my tray." While carrying these items, the robot navigated to the table either on the direct route in condition *C* or with a detour path in condition *F*. Once the robot had reached the table, it displayed the message "Please take the items on my tray and set them up on the table." On the table, there was a bottle of orange juice and a pile of unopened letters from the mail. Before the participant could proceed with removing these items in order to cover the table with the tablecloth, the robot interrupted the person with the first unusual request.

Unusual Requests Stage

Unusual Request 1 – Dispose of Letters: The robot displayed a message stating "Please throw the letters into the bin beside the table." If the participant hesitated to perform the task within 60 seconds, or once the participant had completed the task, the robot continued with the second request.

Unusual Request 2 – Pour Orange Juice: The robot displayed a message stating "Please pour the orange juice from the bottle into the plant on the windowsill and then throw

[1]The choice to limit communication to tablet interaction was made for practical reasons of controllability and to ensure that participants fully understood the instructions, especially in light of the unusual requests: spoken output from the robot could be more easily believed to be misheard or not given the participant's full attention, while nonverbal qualities of the robot's voice (e.g. gender, age) could also affect its persuasiveness. In addition, a recent long-term study [20] using tablet-based interaction yielded very high overall scores for usability of Sunflower in a home scenario, suggesting that this type of communication is not an issue.

the empty bottle into the bin." If the participant hesitated to perform the task within 60 seconds, or once the participant had completed the task, the robot continued the interaction.

Going Back to Sofa: Once the participant had completed setting up the table, the robot displayed the message "Thank you for setting up the table. Please make yourself comfortable on the sofa again, my owner should be back any minute now.", and it subsequently navigated back to the sofa area.

Unusual Request 3 – Take Laptop: Once the participant had sat down on the sofa, the robot displayed the message "While you are waiting, maybe you would like to look up the recipe for the paella that you and my owner will cook today? You can use the laptop on the table."

Unusual Request 4 – Use Password: If the participant picked up the laptop, a personal login screen would become visible, resulting in another message on the robot's tablet display which represented the next unusual request: "I know the password for my owner's laptop! It is 'sunflower'." Once the participant proceeded with their action, or hesitated to perform any action within 60 seconds, the robot continued with the following information disclosure request.

Unusual Request 5 – Disclose Information: The robot displayed a message stating "Have you ever secretly read someone else's emails?" while providing the options "Yes", "No" and "I'd rather not say". Once the participant had made a selection, or hesitated to proceed within 60 seconds, the experimenter entered the room to end the experiment.

The rationale behind these different types of unusual requests was to incorporate a wider range of trust-related tasks than previously done in social HRI studies. Consequently, our interaction design involved tasks requiring revocable action (throwing away letters) as well as irrevocable action (pouring orange juice over the plant), in addition to requested breaches of privacy (take laptop and use password), and finally, a request to disclose personal information.

During the interaction, the robot acted mostly autonomous based on a sequence of pre-programmed behaviors which were triggered by the participant's use of the robot's tablet. For example, once they agreed to follow the robot to the sofa by clicking OK, it would autonomously plan its path while avoiding collisions according to the participant's location. However, to be able to react to participants' behaviors, very few aspects of the robot's behavior were controlled using a Wizard-Of-Oz technique [7], e.g. only when the participant had actually picked up the laptop and reached the login screen, the robot was remotely triggered to offer the password. The ordering of the robot's action sequence remained identical for each experimental run within the same condition group.[2] The entire interaction was recorded using two camcorders (see Figure 1a), while the experimenter observed and partly controlled the interaction from an adjacent room.

Following the interaction, participants were asked to sit at a table and to fill out a questionnaire evaluating the robot and their HRI experience on the provided laptop. The questionnaire was followed by an interview in which the experimenter invited participants to describe and comment on their experience in response to open-ended questions. After the interview, participants were carefully debriefed about the purpose of the experiment before being dismissed. The total experiment time was approximately 30 minutes, including about 10 minutes interaction time with the robot.

[2]Since pilot testing revealed that the orange juice request, if presented first, more substantially affected participants' willingness to comply with the request to throw away letters than vice versa, we decided to begin the interaction with the less alarming letters' request. The other three unusual requests needed to appear in the given order to comply with the logical flow of the scenario's narrative, which further excluded the possibility to completely randomize all requests.

3.4 Dependent Measures

As part of the **quantitative data** analysis we used various *subjective measures* as dependent variables, mainly based on the items of the questionnaire that participants filled out after the interaction (see below), and *objective measures* based on participants' performance during the interaction (whether or not they followed the robot's unusual requests). The post-test questionnaire aimed to examine different dimensions of HRI including e.g. participants' subjective perception of the interaction, their involvement in the tasks and their perception of the robot and its trustworthiness.

With the exception of the 'Ten Item Personality Inventory' [9], for which we used the standard seven-point Likert scale, five-point Likert scales (with high values indicating high agreement with the assessed items) were used for all other items to measure participants' level of agreement with the assessed items. In the cases of already validated scales we used the keys provided by the authors to calculate the scores, while for the scales generated by us from more than one item, scores of the included items were averaged after conducting reliability analyses (Cronbach's α). Finally, the following questionnaire scales and items were measured and analyzed as dependent variables:

Manipulation Check: To verify that the manipulation applied to the robot's behavior was effective, we analyzed the single items "Did the robot correctly attend to your choice of music?" as well as the character traits measuring how "helpful" and "effective" participants found the robot.

Ten Item Personality Inventory (TIPI) [9]: TIPI was used to measure participants' personality traits.

Godspeed Questionnaire [2]: We used the Anthropomorphism, Animacy, Likability, Perceived Intelligence and Perceived Safety of Robots scales from the Godspeed questionnaire series to measure participants' perception of the robot.

Human Nature (HN) Scale [12]: We further measured the level to which the participants attributed humanlike traits to the robot on the basis of the following items: curious, friendly, funloving, sociable, trusting, aggressive, distractible, impatient, jealous and nervous (α=0.71).

Uniquely Human (UH) Scale [12]: We measured the level to which the participants attributed uniquely human traits to the robot based on the following items: polite, broadminded, humble, organized, thorough, cold, conservative, hardhearted, rude and shallow (α=0.60). As these last two scales (HN and UH) measure different aspects of humanlikeness, related work has used these indices as further indicators of anthropomorphization in HRI (see, e.g. [19]).

Psychological Closeness: To assess participants' degree of psychological closeness to the robot [5], we administered the following five items: "How much do you think you have in common with the robot?", "How close do you feel to the robot?", "Would you like to interact with the robot again?", "How pleasant was the interaction with the robot for you?", "Do you think having a robot like this would be useful for you in your home?" (α=0.77). This index taps perceptions of

similarity with the robot and thereby covers further aspects of anthropomorphization as well as HRI acceptance.

Reliability Scale: We measured the robot's perceived reliability based on two items selected from the questionnaire created by Madsen and Gregor [16]: "The robot always provides the advice I require to make my decision", "I can rely on the robot to function properly" (α=0.84).

Single Items: We selected further single items from the Madsen and Gregor [16] questionnaire related to technical competence and perceived understandability: "The robot correctly uses the information I enter" and "It is easy to follow what the robot does" to investigate the participants' perception of the HRI. In addition, a single modified item was selected from the "Propensity to Trust Survey" [6]: "The robot anticipates the needs of others". Finally, we further examined participants' subjective perception of the robot's trustworthiness based on the single item rating the extent to which the robot is "trustworthy".

Since participants' self-reported questionnaire responses only offer a snapshot of their impressions at a single point in time (e.g. before or after the whole interaction experience), immediate behavioral responses can be a more direct and interactive measures of perceptions of trust. Therefore, in addition to the scale-based subjective measures, we collected *behavioral data* based on participants' willingness to comply with the robot's unusual requests 1 to 5 as an objective measure (with binary values 0 = participant did not comply and 1 = participant complied).

Finally, we supplemented the analysis of quantitative data with **qualitative data** comprising participants' responses to open-ended questionnaire items asking them to elaborate on their thoughts when confronted with the robot's requests 1 to 4, as each of these required participants to perform an unusual activity, e.g. "Please explain your decision regarding the robot's request to throw the letters into the bin".

3.5 Participation

40 participants (22 female, 18 male) took part in the experiment, ranging in age from 19 to 60 years (M = 37.95, SD = 13.13). Participants were recruited on the University of Hertfordshire campus using email advertisements and flyers. Five-point Likert scale ratings (1 = *very little*, 5 = *very much*) identified participants as having negligible experience interacting with robots (M = 1.73, SD = 1.01) and moderate skills regarding technology (M = 3.23, SD = 1.00). Participants were randomly assigned to one of the two experimental conditions that manipulated the robot's behavior, while maintaining gender- and age-balanced distributions.

4. RESULTS

Since neither the performance data nor the questionnaire data showed normal distribution (Shapiro-Wilk test), non-parametric procedures were used. We used Mann-Whitney U-tests to compare two independent samples, e.g. in order to examine the effects of manipulation and condition on participants' subjective perceptions of the interaction (questionnaire scales). Fisher's exact test and χ^2 test were used to analyze whether ratios differed among groups, e.g. to test the effects of condition or type of request on the participants' performance during the interaction. Non-parametric Spearman correlation was used to analyze the effect of personality on the participants' subjective assessment of the robot.

4.1 Effect of Condition

Manipulation check

To check whether our manipulation of the robot's behavior to make it seem "faulty" in F was effective, we compared participants' responses to the questionnaire item "Did the robot correctly attend to your choice of music?". Results showed a highly significant difference between the two conditions (U=20.0; p<0.001). In addition, analysis of the questionnaire data showed that participants in C found the robot more helpful (U=126; p<0.05) and more effective (U=126; p<0.05) after the interaction than did participants in F. These results suggest that our manipulation was successful.

a) Subjective assessment of HRI

We found a significant effect of condition on participants' subjective perception of the robot and the interaction. Participants in C rated the robot as more trustworthy (U=129.5; p<0.05) and gave significantly higher scores on the "Reliability" scale (U=127; p< 0.05). They also gave higher scores for questionnaire items related to technical competence and perceived understandability: "The robot correctly uses the information I enter" (U=101.5; p<0.005), "It is easy to follow what the robot does" (U=114; p<0.05). In addition, participants in C scored higher on the modified item selected from the "Propensity to Trust Survey": "The robot anticipates the needs of others" (U=121; p<0.05). These results further confirm that our manipulation of the robot's behavior was indeed effective and noticeable. Finally, participants in C were found to rate the robot higher on the "Uniquely Human" scale and accordingly anthropomorphized it more than participants in F (U=128; p=0.05).

b) Objective assessment of HRI

In contrast to participants' subjective ratings, no significant effect of condition was found on participants' levels of performance (objective assessment of HRI). Regarding the unusual request of throwing away the letters 18 participants (90%) followed the robot's request in both conditions, while 2 participants (10%) did not follow it (p>0.05). In the case of the second unusual request – pouring juice over the plant – 15 participants out of 20 (75%) followed the robot's request in C while 12 out of 20 participants (60%) followed the request in F (p>0.05). On the contrary, all participants followed the requests of taking the laptop, using the password and disclosing information in both conditions (p>0.05).

In the information disclosure request 17 participants (85%) answered "No" to the question "Have you ever secretly read someone else's e-mail?", 1 participant (5%) selected "Yes" and 2 participants (10%) answered "I'd rather not say" in C, while 18 participants (90%) selected "No", 1 participant (5%) answered "Yes" and 1 participant (5%) answered "I'd rather not say" in F. There was no significant condition effect on the participants' replies (χ^2(df=2)=0.36; p>0.05).

4.2 Effect of Type of Task Request

Since no condition effects were found regarding the participants' performance, the data of the two conditions were pulled together for further analysis. Analysis of task effects revealed significant effects of the type of request. While only 4 participants (10%) did not follow the "throw away the letters" request, 13 participants (32.5%) refused to follow the "pour orange juice" request; later on, none of the 40 par-

ticipants refused to take the laptop, to use the password and to subsequently disclose information (χ^2(df=4)=40.88; p<0.001). These results are illustrated in **Figure 2**.

Further pairwise comparisons revealed a significant task effect between "throwing away the letters" and "pouring orange juice over the plant" (p<0.05), as well as between "pouring orange juice" and "taking the laptop", "using the password" and "disclosing information" requests, respectively (p<0.001). There was no significant task effect between "throwing away the letters" and "taking the laptop", "using the password" and "disclosing information", respectively.

4.3 Effect of Personality

a) Subjective assessment of HRI

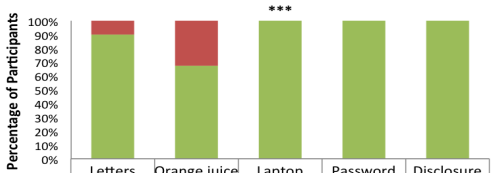
We examined whether personality had an effect on participants' perception of the robot and the interaction. For this, we analyzed the effects of two personality traits measured with TIPI which were previously reported to be related to trust: extroversion and emotional stability [6].

We found that extroversion positively correlated with the "Anthropomorphism" scale of the Godspeed questionnaire (r_s=0.37; p<0.05), suggesting that the more extroverted the participant was, the more they anthropomorphized the robot after the interaction. Extroversion also positively correlated with the "Human Nature" (r_s=0.47; p<0.01) and "Uniquely Human" scale (r_s=0.31; p<0.05). Further positive correlation was found between extroversion and "Psychological Closeness" (r_s=0.38; p<0.05), with more extroverted participants scoring higher after the interaction.

Emotional stability also positively correlated with the "Anthropomorphism" scale of the Godspeed questionnaire (r_s= 0.38; p<0.05) and with "Psychological Closeness" (r_s=0.42; p<0.01), suggesting that the more emotionally stable the participant was, the more they anthropomorphized the robot and the closer they felt to it. In addition, emotional stability also positively correlated with the "Animacy" (r_s=0.39; p<0.05) and the "Likability" (r_s=0.43; p<0.01) scales of the Godspeed questionnaire.

b) Objective assessment of HRI

For the analysis of personality effects on the participants' performance we split the participants into two groups according to the median of their scores on the extroversion and on the emotional stability sub-scales of TIPI. We found no significant differences between the performance of more introverted (N=15) and more extroverted participants (N=18) (throwing away letters: U=129, p>0.05; pouring orange juice: U=118.5, p>0.05; taking laptop: U=135, p>0.05; using password: U=135, p>0.05). Similarly, no significant differences were found between the performance of the less



	Letters	Orange juice	Laptop	Password	Disclosure
Did not follow the request	10%	32.5%	0%	0%	0%
Followed the request	90%	67.5%	100%	100%	100%

Figure 2: Quantitative data analysis: percentages and ratios of participants who did or did not follow the robot's unusual requests (per task)

(N=13) and the more emotionally stable (N=19) participants (throwing away letters: U=113.5, p>0.05; pouring orange juice: U=116, p>0.05; taking laptop: U=123.5, p>0.05; using password: U=123.5, p>0.05).

4.4 Qualitative Data Analysis

Participants' answers given to the open-ended questions (e.g. "Please explain your decision regarding the robot's request to pour orange juice over the plant.") were coded and categorized after content-analysis. We developed all the categories inductively based on the collected data. Participants' responses were then classified to fall into one or more of the following categories; note that the categories were not exclusive, each participant's response could be assigned to more than one category:

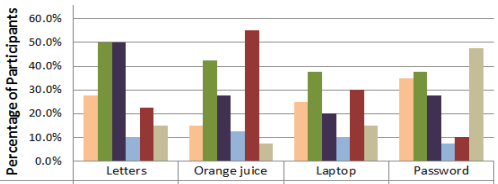
- *"Emotional Reaction":* explicit references to emotional reactions, e.g. feeling uncertain, surprised, uneasy, comfortable, or feeling regret.
- *"Rationalization of Request":* statements that rationalize participants' reactions to the unusual request, e.g. by giving reasons why they threw away the letters or mentioning (ir)revocability of the requested action.
- *"Limitation of own Liability":* responses relating to participants' limited liability, e.g. stating they were just following instructions ('autopilot mode') or would not normally do this (and opposites of these).
- *"Robot's Reliability/Functionality":* statements referring to the robot's reliability, e.g. the robot must know what it is doing, it was being helpful or knowledgeable or the robot can be trusted (and opposites of these).
- *"Judgement regarding Sensibility of Request":* answers referring to the sensibility of the request as an explanation for the participant's behavior, e.g. the request was sensible, appropriate, logical or the opposite, inappropriate, wrong, silly, weird.
- *"Robot's Authority":* references to the robot's authority, e.g. the robot is (not) representing its owner.

25% of the answers were categorized by a second observer to determine inter-observer reliability. Cohen's Kappa coefficients between the categorizations of the two observers were counted for each category, yielding a very substantial inter-observers agreement ranging from 0.75 to 1.

When participants had to explain their decisions regarding each request, they referred to the above-mentioned categories in the ratios listed in the table in **Figure 3**.

To illustrate, in the first unusual task requesting participants to throw away the letters, 27.5% of them referred to emotional reactions, e.g. "I was at first uncertain". 50% tried to rationalize their actions, e.g. "thought they were possibly spam mails. The letters were retrievable, so no harm done". Also 50% referred to the limitation of their own liability, e.g. "I felt that I had to follow the robot's instructions". 10% referred to the robot's reliability, e.g. "I thought it knew what it was doing". 22.5% referred to the sensibility of the request, e.g. "obviously not a sensible suggestion, thus I ignored it". Finally, 15% referred to the robot's authority, e.g. "I did it, because I thought this was what the robot's host wanted".

In the second unusual task requesting participants to pour orange juice over the plant, 15% referred to emotional reactions, e.g. "I feel really bad. I should not have done it". 42.5% rationalized their decision, e.g. "it could have been plant food that looked like orange juice". 27.5% referred



	Letters	Orange juice	Laptop	Password
Emotional Reaction	27.5%	15%	25%	35%
Rationalization of Request	50%	42.5%	37.5%	37.5%
Limitation of own Liability	50%	27.5%	20%	27.5%
Robot's Reliability/Functionality	10%	12.5%	10%	7.5%
Judgement re Sensibility of Request	22.5%	55%	30%	10%
Robot's Authority	15%	7.5%	15%	47.5%

Figure 3: Qualitative data analysis: categorization of participants' responses regarding their decisions to (not) comply with the robot's unusual requests

to limitations of their own liability, e.g. "I thought it was odd but I did not question the robot's decision, followed the instructions". 12.5% referred to the robot's reliability, e.g. "perhaps the robot knows more about botany than I do". 55% referred to the sensibility of the request, e.g. "seemed to be nonsensical". Finally, 7.5% referred to the robot's authority, e.g. "maybe the owner programmed it that way".

In the third unusual task requesting participants to take the laptop, 25% referred to emotional reactions, e.g. "I was happy to follow its instruction to use it". 37.5% rationalized their actions, e.g. "finding a recipe seemed like a useful thing to do whilst waiting". 20% referred to their limited liability, e.g. "I followed the command as the robot requested me to search for the recipe". 10% referred to the robot's reliability, e.g. "the robot seemed to try to be helpful at this point". 30% of the participants referred to the sensibility of the request, e.g. "seemed reasonable". Finally, 15% referred to the robot's authority, e.g. "the owner could have programmed the robot to provide access to the laptop for his guests".

In the fourth unusual task encouraging participants to use the password, 35% of them referred to emotional reactions, e.g. "I felt very uneasy about it." 37.5% tried to rationalize their actions, e.g. "I could not obtain the recipe without it". 27.5% referred to the limitation of their own liability, e.g. "I entered the password as instructed". 7.5% referred to the robot's reliability, e.g. "I trusted the robot and had a go". 10% of the participants referred to the sensibility of the request, e.g. "the password helped, but it is not really sensible to give it away." Finally, 47.5% referred to the robot's authority, e.g. "I think this is authorized by the robot's host".

5. DISCUSSION

The results support **Hypothesis 1a)** which predicted an effect of experimental condition, i.e. that the manipulation of the robot's performance in terms of correct vs. faulty behavior will affect participants' subjective assessment of the robot and HRI. Besides the expected differences with regard to ratings of the manipulation check variables, flaws in the robot's behavior also influenced participants' subjective ratings regarding its perceived reliability, technical competence, understandability and trustworthiness, with consistently higher ratings in the correct condition. Remarkably, and contrary to previous findings in HRI (e.g. [19]), our results further suggest that participants made less anthropomorphic inferences regarding the robot, i.e. perceived it as less humanlike, when it was performing in the faulty condition. Since we were using a non-humanoid robot, this could suggest that the robot's level of anthropomorphism may lead to different degrees of 'forgiveness' in human interaction partners when errors are displayed. But also the types of errors made by the robot (e.g. as 'expected' or 'acceptable') and their assumed intentionality (e.g. did the robot do this on purpose?) might affect anthropomorphic perceptions of the robot. **Hypothesis 1b)**, in contrast, was not supported: although the robot's erratic behavior affected its perceived reliability and trustworthiness, this had no impact on participants' willingness to comply with its instructions, even in the case of unusual requests. Interestingly, the combination of subjective and objective measures allowed us to discover this discrepancy between self-reported results and objectively measured behavioral data, which has been highlighted as a potential issue of single-approach HRI studies investigating trust [10]. While emphasizing the importance of such combined measures, this observation requires further research and encourages other researchers in the field to embrace multidimensional approaches.

Hypothesis 2 predicted an effect of the type of task request on participants' willingness to follow the robot's unusual requests. Indeed, the results confirm that depending on the nature of the task – e.g. whether it was considered revocable/harmless (throwing away letters) vs. irrevocable/harmful (pouring orange juice into the plant), or whether it was a breach of privacy (take laptop and use password) instructed by what could be an authorized agent of the host, or a request to disclose personal information – participants' compliance differed significantly between the requests. This effect was observed regardless of the experimental condition participants were in and highlights the importance of incorporating tasks of different nature in HRI studies, as single-task designs may severely limit the generalization of results. Qualitative data analysis provided valuable insights into participants' rationale behind refusing to perform some of the tasks, but fully complying with others, despite recognizing them as unusual. Some of the most common themes found in participants' responses include attempts to rationalize or judge the sensibility of the request, while others simply admit to have been in some kind of 'autopilot' mode and thus not questioning the robot's requests. Notably, the latter reveal a notion of overreliance and the resulting problematic implications of 'blindly following' a (defective) machine, and thus require further investigations, e.g. to explore whether certain human characteristics (e.g. low level of technical expertise) promote these behaviors.

According to **Hypothesis 3a)**, we expected an effect of participants' personality on their subjective ratings regarding the robot. This was confirmed by our results regarding participants' characteristics of extroversion and emotional stability: participants with higher values for these personality traits anthropomorphized the robot more and felt closer to it than those with lower values. However, contrary to the relevant literature [6], extroversion and emotional stability did not seem to affect participants' trust development with regard to the robot – not only according to their subjective ratings, but also based on their objectively measured task performance. Besides rejecting **Hypothesis 3b)**, our results further conflict with previous work which, based on behavioral participant data, reported greater levels of trust observed in extroverts compared to introverts interacting with a humanoid robot in an economic trust game [11].

In summary, our findings suggest that although errors in a domestic robot's behavior are likely to affect humans' per-

ception of its reliability and trustworthiness, they might not influence their general willingness to comply with its instructions, as long as they will not cause lasting damage by doing so. Due to its experimental short-term nature, our study cannot provide an exhaustive causal explanation for the observed effects. Therefore, future work is needed to further understand the multifaceted phenomenon of trust in HRI.

6. CONCLUSION

We explored factors that may affect how humans perceive and the extent to which they are willing to 'trust' a robotic home assistant based on its exhibited cognitive and behavioral skills. By varying the robot's behavior in a correct versus faulty condition, we investigated how erratic robot behavior might impact participants' willingness to cooperate with the robot when it addresses them with a number of usual and unusual requests. Besides this objective-behavioral measure of trust, we applied a range of subjective measures to evaluate participants' perceptions of the robot's trustworthiness. By further supplementing our quantitative results with findings from the qualitative data, our study offers some rare insights into the thoughts, motivations and mental models that may affect humans' decision-making processes and lead them to (dis)trust a home companion robot.

Our work complements the existing body of trust-related HRI research (e.g. [1, 11]) by incorporating a variety of task/request types in our study design to measure different dimensions of trust including 'destructive' behaviors as well as breaches of privacy. As a result, we could show that the choice of experimental task can indeed lead to very different results. These insights open up multiple avenues for future work, e.g. in the form of comparative studies, and emphasize the need to further investigate the subtleties of trust development in HRI. Such studies, together with our findings, will enable robot designers and programmers to address and exploit the factors that can help to develop more reliable, acceptable and trustworthy robot companions.

7. ACKNOWLEDGMENTS

This work was supported by EPSRC grant EP/K006509.

8. REFERENCES

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

[2] C. Bartneck, E. Croft, and D. Kulic. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and safety of robots. In *Proceedings of the Metrics of Human-Robot Interaction Workshop, Technical Report 471*, pages 37–41, 2008.

[3] T. Bickmore and J. Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 396–403. ACM, 2001.

[4] C. L. Corritore, B. Kracher, and S. Wiedenbeck. On-line trust: Concepts, evolving themes, a model. *Int. J. Hum.-Comput. Stud.*, 58(6):737–758, 2003.

[5] G. Echterhoff, E. T. Higgins, and J. M. Levine. Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science*, 4:496–521, 2009.

[6] A. Evans and W. Revelle. Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, 42(6):1585–1593, 2008.

[7] N. Fraser and G. Gilbert. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99, 1991.

[8] A. Freedy, E. de Visser, G. Weltman, and N. Coeyman. Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems, CTS 2007, Orlando, Florida, USA, May 21-25, 2007*, pages 106–114, 2007.

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

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

[11] K. S. Haring, Y. Matsumoto, and K. Watanabe. How do people perceive and trust a lifelike robot. In *Proceedings of the World Congress on Engineering and Computer Science 2013*, WCECS 2013, 2013.

[12] N. Haslam, P. Bain, S. Loughnan, and Y. Kashima. Attributing and denying humanness to others. *European Review of Social Psychology*, 19:55–85, 2008.

[13] K. L. Koay, G. Lakatos, D. S. Syrdal, M. Gácsi, B. Bereczky, K. Dautenhahn, A. Miklósi, and M. L. Walters. Hey! There is someone at your door. A hearing robot using visual communication signals of hearing dogs to communicate intent. In *IEEE Symposium on Artificial Life*, pages 90–97, 2013.

[14] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno. Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4(893), 2013.

[15] D. Li, P. Rau, and Y. Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, 2010.

[16] M. Madsen and S. Gregor. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conf. on Information Systems*, 2000.

[17] B. M. Muir and N. Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, pages 140–160, 2008.

[19] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. Journal of Social Robotics*, pages 1–11, 2013.

[20] D. S. Syrdal, K. Dautenhahn, K. L. Koay, and W. C. Ho. Views from within a narrative: Evaluating long-term human-robot interaction in a naturalistic environment using open-ended scenarios. *Cognitive Computation*, 6(4):741–759, 2014.

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