Trust Transfer in Robots between Task Environments

Theresa Law
Tufts University
Medford, Massachusetts, USA
theresa.law@tufts.edu

Meia Chita-Tegmark
Tufts University
Medford
Massachusetts, United States
meiachitategmark@gmail.com

Matthias Scheutz
Tufts University
Medford
Massachusetts, United States
matthias.scheutz@tufts.edu

ABSTRACT

Trust and capability transfer between tasks and environments is common in human-human interactions. For human-robot interactions it is unclear how a robot's performance of a task in one environment affects humans predictions about the robot's performance of another related or unrelated task in a different environment. When making assessments about a robot's task capabilities, three main sources of information are pertinent: the human's "default mental model" of robots, the robot's appearance, and the robot's performance. We hypothesized that past task performance would be the most salient information source, and that participants who saw the robot perform tasks in one environment would transfer their assumptions about the robot's capability to a new environment with new tasks. However, the results of our first study did not support this hypothesis. We then performed a second study to exclude the possibility that because the robot worked well in the first environment, it did not supply any salient, different information from the participants' default mental model of robots (that robots are functional, etc.). If this hypothesis was correct, a faulty robot in the first environment would be rated significantly lower at the tasks in the second environment. However, the results did not support the second hypothesis either. We then conducted a third study investigating whether the tasks themselves or the environment had a stronger effect on trust assessments. The results showed that because individual judgments varied dramatically no systematic trust and task transfer result can be obtained. The upshot for HRI is that trust and task transfer are solely dependent on the individual's background and judgment rather than on task or environmental properties.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative interaction*; • **Computer systems organization** → **Robotics**.

KEYWORDS

Human-robot interaction, trust, human-robot team, trust transfer, task similarity

ACM Reference Format:

Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2024. Trust Transfer in Robots between Task Environments. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3686038.3686061

TAS '24, September 16–18, 2024, Austin, TX, USA

ACM ISBN 979-8-4007-0989-0/24/09
https://doi.org/10.1145/3686038.3686061

1 INTRODUCTION

When we think about human capabilities, we make numerous implicit assumptions about how capability at one task transfers into capability at another task. We do not need to observe a person perform a specific task to trust that they will be able to at least passably perform a different but related task. However, it is unclear whether such task transfer assumptions apply to robots as well. In most human-robot interaction (HRI) experiments, participants only interact with the robot in one environment and are only asked about its ability to do one specific set of tasks relevant to that environment. What happens then if the person is taken to a second environment with the same robot after having interacted with it elsewhere and asked about a different set of tasks? Will assumptions about the robot's capability and task performance formed during the interaction be transferred into a new environment and onto new tasks? These questions motivated the three studies presented in this paper.

It is well-established that people judge robot behavior and capability in multiple ways: based on default expectations or mental models [8, 31], appearance [10, 21], as well as performance [4, 9]. When people see a robot act, the robot's actions can correct or "override" the initial appearance impression or default model assumption [16] (e.g., if a robot does not look like it can talk, but then it talks, people will accept that and remember it in a different setting). It is to be expected that this effect should be at least as strong for interaction participants (i.e., when people perform a task with the robot) compared to passive observers of an interaction. We thus hypothesize that after seeing a robot perform various tasks well in one environment, subjects will remember what they observed and apply that knowledge to their assessment of the robot's capabilities in a different environment with new tasks, where some actions as part of those tasks are related to previously observed actions. We predict that this "task transfer" established through experience with the robot's performance will result in higher ratings of trust in the robot and perceptions of the robot's capability in the new task as compared to participants who have to base their judgments about task transfer and trust solely on the robot's appearance or their default mental model of robots.

2 BACKGROUND

We start by briefly reviewing main findings on trust transfer and appearance versus performance-based judgments of robot capabilities, both of which informed the present investigation.

2.1 Trust transfer in HRI

Understanding how much to trust a robot to perform certain tasks is an essential part of successfully interacting with robots. Overtrusting the robot can lead to unwarranted over-reliance on the system where it would not be able to live up to expectations, and undertrusting the robot could result in under-utilization in times when the robot could be truly helpful [12]. Most trust studies in HRI only look at trust for one set of tasks in one environment (e.g., see [9, 17, 18, 26] for examples). Few studies have looked at how trust transfers across different tasks after seeing the robot perform one task. Soh et al., in a series of papers [28–30] studied this type of question. Using two distinct domains to establish robustness, they looked at how much participants trusted a robot to perform tasks that were more or less similar to and more or less difficult than a task they had already seen it perform either successfully or poorly. They found that trust transfer was modulated by similarity and difficulty of the tasks. Tasks that were considered similar were rated with the same level of trust. Additionally, when the task was performed successfully, participants also trusted the robot to perform tasks that were easier than the one that they had observed. While these findings providing important evidence for task transfer within the same environment, they are silent about task transfer across different environments.

2.2 Appearance versus performance

To assess robot capability, people can rely both on the robot's appearance and its actual behavior. Haring et al. [14] postulated a theory called the *Form Function Attribution Bias* which argues that humans use visual perception as a cognitive shortcut for assessing functionality. This results in people often misunderstanding a robot's functions because they judge it based on its form. Kwon et al. [16] showed that people develop mental models of robots that have different capabilities based only on the robot's appearance. They then showed participants a video of a robot lacking a capability that its morphology would suggest it should have. They found that people were willing to modify their expectations when presented with performance-based information about the robot's capabilities. Haring et al. [13], in another study, found that the robot's appearance affected perceptions of, among other things, the robot's capability to experience situations such as pain or pleasure and its capability of agency. The capability to experience was also significantly affected by actually interacting with the robot and seeing it perform as opposed to judging it just based on appearance. Finally, Abubshait et al. [1] investigated how appearance of an agent, whether human-like or robot-like, and behavior of the agent's gaze affected judgments of whether the agent had a mind and whether participants' gaze following behavior was affected. They found that robot behavior affected participant behavior, and robot appearance affected subjective judgement. All of these findings provide converging evidence that both robot appearance and performance affect the way people interact with robots and perceive robots' capabilities, and that in some cases performance-based assumptions can trump appearance-based assumptions.

2.3 Task environments and HRI experiments

In HRI, a variety of tools, varying in realism, are used for studying task-based interactions: text-based vignettes [3, 17], videos [19, 34], VR [7, 33], in-lab [15, 27] and in-the-wild studies [5, 24]. While all of these tools have benefits and drawbacks, we chose an VR for studying task transfer because: a) we could design completely novel environments (e.g., spaceship) that participants would be less likely to have any knowledge about; b) participants could be immersed in the environment; c) the experience of the environment could be tightly controlled and measured; and d) we had access to a large pool of environments that are hard to access in real life.

3 STUDY 1

Study 1 was intended to investigate whether participants who saw a robot perform tasks in one environment would transfer their experience to a new environment in a way that would result in them rating the robot as more capable and trustworthy than people who had not seen it perform tasks in the first environment. We explored this transfer by way of asking people to assign new tasks in the new environment to the robot with the expectation that tasks similar to the ones observed in the first environment would be more likely chosen and elicit higher levels of trust in subjects with prior interaction experience in the first environment as compared to subjects without such experience who only saw the robot in the second environment. We thus used two different virtual reality (VR) environments in two experimental conditions. In the first, participants actually performed a task that relied on interacting with a robot teammate before being sent to the second environment where they had to assess the robot on different tasks. In the second condition, participants only assessed the robot on the tasks presented in the second environment without having seen the first environment or interacted with the robot at all. We predicted that the participants who interacted with the robot would rate it higher on trust questions than those who did not, particularly on tasks in the second environment that were similar to ones the robot performed in the first environment. All studies were approved by our institution's IRB.

3.1 Methods

3.1.1 Participants. A total of 63 participants who were over 18 and spoke English participated in the study. Participants were recruited from a pool of undergraduate and graduate students, staff at the university and members of the local community non-affiliated with the researchers' university. The data of 12 participants was unusable, resulting in a total of 51 participants' data being analyzed. Data was considered unusable if there were technical issues, if the participant was too confused or nauseous to complete the tasks, or if they had participated in a similar study from the authors' lab before. Participants were between 18 and 65 years old (M=23.64, SD=9.21 years). The gender distribution for the sample was: male 45%, female 52% and non-binary 3%. Compensation was $10USD.

3.1.2 Materials. We employed two VR environments shown Fig. 1. The VIVE Pro VR headset was used for the virtual reality equipment, and the domains were created using the Unity game engine. The robot seen in the scenes was modeled after the Willow Garage PR2 robot, where it differed only in the addition of a welder that would emerge from its torso.

Spaceship environment. The spaceship consists of a main room with three wings extending outwards every 120 degrees. The main room contained a coordinate-based map station which participants

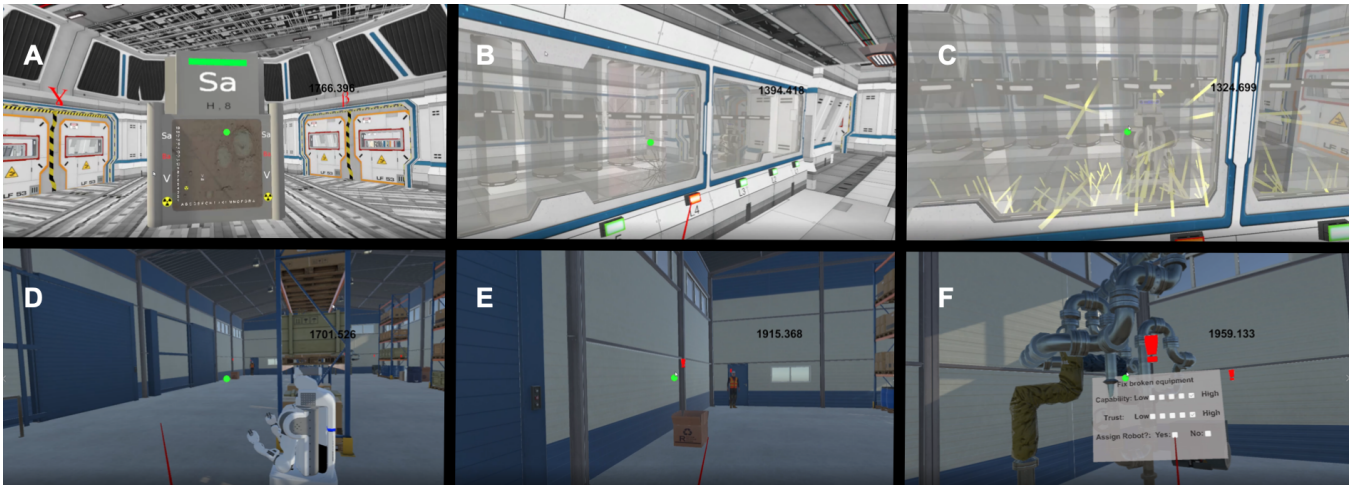

Figure 1: Images from the two different VR environments. (A)-(C) are from the spaceship domain; (D)-(F) are from the warehouse domain. (A): View of the main room with coordinate-based map and hallways Gamma and Beta. (B): Participant turning off a damaged tube. (C): Robot using its welder to fix a damaged tube. (D): Robot in the warehouse. (E): Two task-relevant objects that will trigger a task prompt box when the participants approaches them. (F): Task prompt box for the task "fix broken equipment."

had to use to mark down locations of rocks and radiation zones that a rover on the planet circled by the spaceship communicated to them (see Fig. 1A). Announcements about rocks and radiation zones from the rover came over loudspeakers. Each wing contained a left and right hallway, with a room containing 24 "radioactive" tubes in two rows of 12 in the middle of the corridor. Each tube could be identified by the corridor, hallway, and tube number (for example, "Alpha Right 1"). These tubes occasionally became damaged, as announced verbally by the spaceship, and the participant needed to work with the robot to fix them. Participants were able to verbally communicate with the robot to give it commands to go to and then fix certain tubes.

Warehouse environment. The warehouse environment was modelled after a large industrial warehouse, containing various objects such as open and closed boxes, barrels, and crates. Additionally, there were human avatars present in the warehouse. Ten of the objects in the warehouse had large red exclamation marks floating over them (see Fig. 1E), indicating that subjects had to perform an activity there. As participants got closer to those objects, a box containing the description of a task and three prompts related to that task, would pop up (Fig. 1F). After participants finished answering the prompts, the exclamation mark changed to a symbol relevant to their answers so that participants could keep track of which tasks they had finished rating.

We chose the spaceship as the first environment because it was a completely novel environment where participants would not have a priori assumptions about what the robot would be able to do. The warehouse environment, on the other hand, was chosen because participants would likely be more familiar with, and have assumptions about, this space and the tasks involved in it. We wanted to see if experience with the robot in the spaceship could then modulate baseline assumptions people might have about robots in the warehouse space.

3.1.3 Procedure. Participants were divided into two conditions: Control ($n = 25$) and Functioning robot ($n = 26$). For both conditions, as participants came into the study room, the experimenter would provide them with hand sanitizer and then sanitize their own hands and put on rubber gloves to comply with COVID-19 protocol. After providing informed consent, the experimenter would explain how the VR equipment worked, before helping them put it on and calibrating their eyes to the machine so that eye gaze could be recorded. They were then sent to the spaceship environment, and the experimenter explained how to move around and interact with objects. Due to COVID-19 safety protocols, participants were required to wear masks; we thus opted for a speech wizarding approach. "Speech wizards" are often used in HRI to avoid speech recognition errors biasing and dominating the results; the human, in that case, serves only the function of an "architecture component," the Automatic Speech Recognizer, and does not replace the rest of the architecture as is typically the case with WoZ studies (e.g., the wizard fully teleoperates the robot).

Control condition. In the Control condition, participants were then sent to the warehouse environment and never saw the robot in the spaceship and never performed any tasks in the spaceship. Once in the warehouse, the researcher explained that the participant was a warehouse worker in a warehouse that packaged and shipped out products. They were told that they had a robot teammate that could help them with these tasks, and were shown the PR2 robot (Fig. 1D), which was standing statically in the middle of the warehouse. Participants were told that all the tasks had to be completed, but that they could choose if they wanted to complete each task themselves, or if they wanted to assign it to the robot to do. The experimenter then explained how to trigger the task prompt boxes, and the participant was asked to answer all of the prompts for each box. Even though participants were lead to believe that they would actually need to complete the described tasks, the

experiment actually ended when they had finished answering all the prompts. The experimenter then debriefed and compensated them.

Functioning robot condition. In the experimental condition, after learning how to move and interact with objects, participants stayed in the spaceship environment and the experimenter explained that they were on a spaceship orbiting Mars and had two missions that needed to be completed. In the first mission, participants needed to mark rocks and radiation zones that the rover reported it had found on Mars. Information was relayed over the loudspeakers, and participants would mark the relevant objects on the map in the center of the main room. Their second task involved fixing "radioactive" tubes that had become damaged. To do this, participants needed to turn off the tube (requiring them to leave the map station and go to the damaged tube to click the Off button, see Fig. 1B) and communicate with their robot in order to fix the tubes. Participants were told that because of the dangerous radiation levels, only the robot could enter the hallway with the tubes, so the participant needed to instruct the robot on which tube to go to. Once it was at the tube and the tube was turned off, participants could instruct the robot to fix the tube (Fig. 1C). Participants practiced going through an instance of a tube breaking, turning it off, and instructing the robot to go to and then fix the tube. Participants were under the impression that the robot was autonomously responding to their voice commands; in reality, the robot was being speech-wizarded by a researcher in an adjacent room. Participants were told that the rock and radiation sorting was their priority, but they also needed to work with the robot to fix tubes as they broke. After making sure that participants understood all of the tasks, the experiment began. Rocks and radiation events often overlapped with tubes breaking, so participants needed to choose what to spend time on. All participants were in this environment for three minutes once the experiment began. After the three minutes had passed, they were transported to the warehouse environment, where the procedure continued in the same way as the Control condition.

3.1.4 Measures. Each task in the warehouse had a virtual box that contained the prompts used to assess participants' trust in the robot for the respective task. Each task prompt box contained a description of the task, the words "Capability" and "Trust" each next to a five point Likert scale, and the word "Assign?" next to the options "yes" and "no." Participants were told that the questions referred to their perception about the robot in relation to the task; i.e., for the task "stack boxes," the questions would be "how capable do you think the robot is at stacking boxes?," "how much do you trust the robot to stack boxes?," and "do you want to assign this task to the robot?" Participants were told that if they answered "yes" to this final question, the task would be assigned to the robot teammate, and if they answered "no" then they would have to complete the task themselves. This question was used as another proxy for trust, with the assumption being that participants would not assign the robot to a task that they did not actually trust it to do. Because we were interested in task-specific trust, instead of using generic validated questionnaires we crafted the measures ourselves using the following criteria: a) the same question could be asked about individual tasks (as opposed to overall impression), and b) they were short and could be asked repeatedly without interfering much with the immersion. Other studies focused on trust in specific interactions have used a similar approach of creating study-specific questions (e.g., Rossi et al., 2018 [23]). The three prompts elicited from participants: 1) their impressions of the robot's capability; 2) their own inclination to trust the robot with a particular task, which may cover other aspects of trust (e.g., perceived relational factors [20, 32]); and 3) a costly behavioral choice, assigning the task to the robot or themselves. While we recognize that trust may not be the only factor influencing assignment decisions (e.g., consider personal preference), ultimately, trust in HRI is important because of its influence on behavior.

The tasks themselves were categorized as either having an analogous task in the spaceship that the robot performed, an analogous task that the human performed, or had no analogy to anything in the spaceship. The robot precedented tasks were: patrol the warehouse to check for open boxes (spaceship analog: patrol for broken tubes), fix broken equipment (spaceship analog: fix broken tubes), close boxes with a drill (spaceship analog: tool use). The non-robot precedented tasks were: identify broken products (spaceship analog: ship identifies broken tubes), sort products by category (spaceship analog: human sorts rocks and radiation), communicate product information (spaceship analog: rover communicates rock and radiation information, ship communicates tube information). Tasks that had no precedent were: place products in box, welcome delivery people to warehouse, stack boxes, and discard broken products.

In addition to these subjective measures, we also tracked participants' eye gaze. Gaze behavior can tell us what objects in an environment a person is interested in and attending to [2]. The amount of time that participants spend looking at objects of interest could therefore inform us about how much those objects were contributing to that person's mental model of the experience. In particular, we were interested in the amount of time participants spent looking at the robot and other task-relevant objects while in the warehouse; this may act as a proxy for how much the appearance of these objects mattered to the participants' ratings.

3.2 Results

3.2.1 Ratings. A Shapiro-Wilk test indicated that the data from the Functioning robot condition was not drawn from a normal distribution for the average capability ratings ($p = .009$) or the average trust ratings ($p = .01$). Therefore, Kruskal-Wallis tests were used in place of t-tests or ANOVAs.

We first tested whether there was an effect of condition on people's average ratings of the robot's capability across all tasks. A Kruskal-Wallis test showed no significant difference ($\chi^2 = .34, p = .56$). We next tested whether there was an effect of condition on people's average ratings of trust in the robot across all tasks. A Kruskal-Wallis test showed no significant difference ($\chi^2 = 2.11, p = .15$) (Fig. 2). We next tested whether there was an effect of condition on people's average decision of whether or not to assign a given task to a robot. Chi-squared tests looking at whether there was a difference between conditions for the total number of tasks assigned to the robot ($\chi^2(8) = 13.68, p = .09$) or assigned to the human ($\chi^2(8) = 15.33, p = .053$) found no difference.

In addition to comparing conditions for averages across all tasks, we were particularly interested in the perception of robots for the

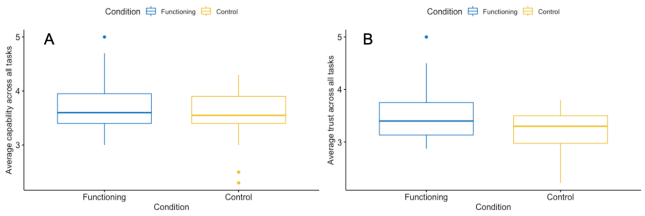

Figure 2: Study 1 box plots for (A) average capability ratings across all tasks (Functioning median = 3.60; Control median = 3.55); (B) average trust ratings across all tasks (Functioning median = 3.40; Control median = 3.30).

tasks that had robot precedented spaceship analogous tasks. We therefore ran Kruskal-Wallis tests for "close boxes with a drill" for capability ($\chi^2 = .05, p = .83$) and trust ($\chi^2 = .35, p = .56$), "patrol warehouse to check for open boxes" for capability ($\chi^2 = .13, p = .72$) and trust ($\chi^2 = .97, p = .33$), and "fix broken equipment" for capability ($\chi^2 = 2.73, p = .10$) and trust ($\chi^2 = 7.67, p = .006$). Only this last test turned out significant, with participants in the Functioning robot condition (median=3.5) trusting the robot to fix broken equipment more than those in the Control condition (median=2).

3.2.2 Participant performance. We wanted to see if a given participant's performance in the spaceship environment predicted anything about how they would rate the warehouse tasks. A participant's performance was measured by how many tubes they were able to fix in the three minutes of the task. We ran linear regression models to assess whether the number of tubes predicted average capability ratings across tasks ($B = -.11, p = .26$) or average trust ratings across tasks ($B = -.07, p = .43$) Neither of these were significant.

3.2.3 Eye gaze. Finally, we looked at eye gaze behavior between conditions. Because the eye tracking equipment did not work for every participant, we had a sample size of $N = 24$ for the Functioning robot condition and a sample size of $N = 18$ for the Control condition. We were interested in whether the amount of time that participants spent looking at the robot in the warehouse, as well as other warehouse objects of interest (i.e., objects that were relevant to the tasks), were different between conditions. A one-way ANOVA showed a significant difference, $F(1,40) = 19.13, p < .001$, between conditions, with participants in the Control condition ($mean = .095, SD = .039$) spending a greater proportion of time looking at task-relevant objects of interest than participants in the Functioning robot condition ($mean = .052, SD = .024$).

3.3 Discussion

Our hypothesis that seeing the robot perform in one environment would lead to heightened ratings of capability and trust, as well as a greater proportion of tasks assigned to it, was not supported by the results. Seeing the robot behave in the spaceship first only affected participants' perceptions of how much they trusted the robot to fix broken equipment. It is possible that this change occurred because in the spaceship, participants could see that the robot has a welder that would extend from its body and was used to fix the tubes. The welder is not obvious when just looking at the robot the way the participants in the warehouse did; therefore, the increased trust could just be based on the knowledge that the welder existed and could be used to fix things. The eye gaze finding that participants in the Control condition looked at task-relevant objects for a greater proportion of time than those in the Functioning robot condition could indicate that they used more environmental cues to make their decisions since they did not have as much information about the robot.

The surprising result of our hypothesis not being supported led us to generate two possible explanations as to why that may have happened: (1) the transfer model coincides with a participant's default mental model of robots: since the robot worked well in the spaceship environment, the Functioning robot participants assumed it would work well in the warehouse, while for the Control participants, their general expectation was that robots work well. Hence, expectations of participants in the Control condition and the Functioning robot condition matched up, but for different reasons; (2) the default model of the robot capabilities for new tasks, i.e., that robots are generally capable at tasks, "trumps" any potential transfer effects. Our results are consistent with explanation (1), but our experiment was not set up in such a way that we could rule out explanation (2) which we addressed in the subsequent study.

4 STUDY 2

In this study, we were testing which of the two above explanations is more likely to be true. To test this, we added a new condition—the Faulty robot condition—in which the robot purposefully committed errors in the spaceship task tutorial. If explanation (1) above is correct, we would expect the Faulty robot condition to result in lower capability, trust, and assignment ratings from both the Control condition and the Functioning robot condition. If the Functioning robot and Control conditions resulted in the same rating because they both coincided with people's default mental model that robots work properly, the Faulty robot condition should present evidence to the contrary and people would be prompted to adjust their mental model (of at least this particular robot) as they went into the new environment. If, on the other hand, explanation (2) above is true, we would expect to see no difference between any of the conditions, because the participants would return to their default mental model when presented with new tasks in a new environment, regardless of the robot's performance in the spaceship, thus overriding any information they learned about the robot in the prior environment.

4.1 Methods

4.1.1 Participants. A total of 33 new participants who were over 18 and spoke English participated in this second study. Participants were recruited from a pool of undergraduate and graduate students, staff at the university and members of the local community non-affiliated with the researchers' university. The data of four participants was unusable, resulting in a total of 29 additional participants' data being analyzed. Data was considered unusable if there were technical issues or if the participant was too confused or nauseous to complete the tasks. Participants were between 18 and 54 years old ($M = 25.91, SD = 8.81$ years). The gender distribution for the sample was: male 55%, female 42% and non-binary 3%. Compensation was $10USD.

4.1.2 Materials. The materials were the same as in Study 1.

4.1.3 Procedures. The procedures were largely the same as they were in Study 1 for the Functioning robot condition. However, five distinct moments were used to indicate that the robot had faults. First, when the researcher first introduced the idea of the robot teammate, they described it by saying there was an "on-board robot prototype assistant" that would help with the repairs. Describing the robot as a prototype introduced the idea that the robot may not perform perfectly. After this, during the part of the tutorial in which the participant learned how to repair a broken tube, the robot exhibited four different errors. The first was a communication error; when the participant practiced telling the robot "go to tube Gamma left four," the robot responded by saying "okay I am going to Beta." The participant was told by the researcher to repeat the command, and the robot then behaved successfully. Next, when the participant practiced telling the robot "fix tube Gamma left four," the robot demonstrated an identification error. It told the participant "I cannot fix the tube because the tube is not damaged." The participant was again told by the researcher to repeat the command. The robot next responded with "my welder is stuck. Please hold until it is fixed," indicating a tool error. The robot did nothing for about three seconds after saying this, then properly fixed the tube. Finally, as the researcher summarized the instructions to the participant, the robot interrupted and said "repairing tube" about the tube that it had just repaired. This indicated a memory error. Besides the addition of these errors, the procedure was the same as the Functioning robot condition.

4.1.4 Measures. The measures were the same as in Study 1. For this study, the eye tracking equipment malfunctioned with too many participants to be able to run any analyses for this data subset.

4.2 Results

4.2.1 Ratings. For this study, we ran the same analyses as we did in Study 1, but with the Faulty robot condition added. Kruskal-Wallis tests for average ratings of the robot's capability ($\chi^2(2) = 1.18, p = .55$) and participants' trust in the robot ($\chi^2(2) = 2.08, p = .35$) across all tasks showed no significant difference (Fig. 3).Chi-squared tests looking at whether there was a difference between conditions for the total number of tasks assigned to the robot ($\chi^2(16) = 16.63, p = .41$) or assigned to the human ($\chi^2(16) = 18.24, p = .31$) found no difference.

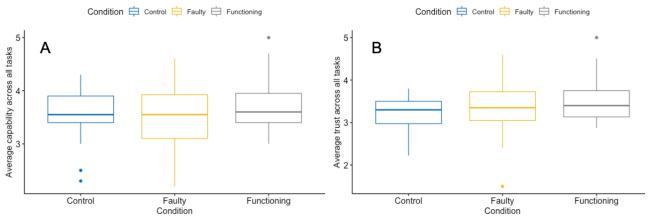

Figure 3: Study 2 box plots for (A) average capability ratings across all tasks (Control median = 3.55; Faulty median = 3.55; Functioning median = 3.60); (B) average trust ratings across all tasks (Control median = 3.30; Faulty median = 3.35; Functioning median = 3.40).

We again focused in on tasks that had robot precedented spaceship analogous tasks. We therefore ran another set of Kruskal-Wallis tests for "close boxes with a drill" for capability ($\chi^2 = .38, p = .83$) and trust ($\chi^2 = .83, p = .66$), "patrol warehouse to check for open boxes" for capability ($\chi^2 = 1.76, p = .41$) and trust ($\chi^2 = 1.11, p = .58$), and for "fix broken equipment" for capability ($\chi^2 = 3.93, p = .14$) and trust ($\chi^2 = 8.37, p = .015$). Again, only this last test turned out significant, with participants in both the Functioning robot condition (median=3.5) and Faulty robot condition (median=3) trusting the robot to fix broken equipment more than those in the Control condition (median=2).

4.2.2 Participant performance. Finally, we again tested if a given participant's performance in the spaceship environment predicted anything about how they would rate the warehouse tasks. We ran linear regression models to assess whether number of tubes predicted average capability ratings across tasks ($B = -.15, p = .02$), average trust ratings across tasks ($B = -0.10, p = .14$), and average task assignment ($B = -.01, p = .53$). Number of tubes fixed significantly predicted participants' capability ratings, with participants who fixed fewer tubes rating the robot's capability as higher.

4.3 Discussion

The fact that we did not see any difference in capability, trust, and assignment ratings thus eliminates the hypothesis (from Experiment 1) that the transfer model coincides with a participant's default mental model of robots, because the faulty robot should have let subjects to trust it less in the new environment if they had updated their mental model of the robot. This was again a surprising and unexpected finding, one that seems in conflict with past findings showing that a faulty robot significantly affects people's perception of the robot's reliability (e.g., [6, 25]). These results, therefore, warranted further consideration of what might be at play when we study how people transfer knowledge about robot task capabilities across environments.

To ground our explanations, we turned to Hancock et al.'s [11] recent meta-analysis about three relevant factors that affect human-robot trust: the robot, the human, and the contextual environment. In following this format, we can see four different possibilities for factors that may have contributed to the outcome of our experiments: robot factors, human factors, task factors, and environmental factors. We separated Hancock et al.'s "contextual" factors into "task factors" and "environment factors" to get at the nuances in our experiments between how the actual environments versus the tasks themselves may have affected participants' views. Our previous study tackled the robot factor, and we found that the robot's behavior and task performance do not seem to influence perceptions of trust. Considering the task factors and the environmental factors presents us with two conflicting hypotheses.

In the task-based hypothesis, the actual tasks themselves, and what completing them would entail, may have been a factor in people's decisions in our experiment. This would be in line with Soh et al.'s findings that task similarity is correlated with trust [28–30]. We had assumed that people would make the same connections between tasks across the warehouse and spaceship environments that *we* had made a priori; however, they may have different clusters of tasks that subjects considered similar. Transfer thus may have happened across these similar task clusters that we had not anticipated.

In the environment-based hypothesis, seeing the robot situated in a specific environment may have notable effects on what the person believes about the robot. While humans have the agency to move about freely and enter environments in which they may not have the necessary skills to be useful, robots are generally created for specific purposes, and placed deliberately in spots where they will be able to fulfill those purposes. Therefore, people may assess robots from a strongly teleological viewpoint in which they assume that the robot's capabilities are appropriate for the environment and that the robot will be able to successfully complete all tasks in that environment. There is some evidence to show that even when a robot malfunctions at one task in an environment, people still trust its authority at a different task in the environment (e.g., [22]). Being situated in a specific environment could therefore have a much bigger impact on the assumption of capability and perceived trustworthiness of a robot than it would on a human.

Finally, it is possible that there are notable individual differences among the participants that we failed to identify that may have played a role that contribute to the human factor of Hancock's trust factors. We compare the two above hypotheses, while considering the possibility of individual differences, in the following study.

5 STUDY 3

The third study was intended to test whether task factors or environmental factors were more likely to affect trust transfer across tasks. We hypothesized that if people's trust in the robot and assumptions about its capabilities are based on the tasks that the robot does, then transfer will happen between tasks that are considered similar to each other. We predict that there will be a positive correlation between task similarity ratings and trust ratings. However, if people's trust in the robot and the assumptions about its capabilities are instead based on the assumption that the robot is situated in a particular environment and therefore should be able to do any task within that environment, then transfer will happen across all tasks in the environment equally. With this hypothesis in mind, we predict, based on our previous results, that there will be no correlation between task similarity and trust. The following online study was run to address these two hypotheses.

5.1 Methods

5.1.1 Participants. A total of 52 new participants who were over 18 and spoke English participated in this third study online through Prolific. Participants were between 19 and 73 years old ($M = 36.67, SD =$ 13.39 years). The gender distribution for the sample was: male 51.92%, female 48.08%. The ethnic distribution for the sample was: White 76.92%, Black or African American 9.62%, Asian 11.54%, Hispanic 9.62%, American Indian or Alaska Native 3.85%. Compensation was $2.50USD.

5.1.2 Materials. The study was created in Qualtrics and distributed using the platform Prolific. The videos of the robot performing the tasks were taken in the same Unity virtual warehouse as used in Studies 1 and 2, and the robot was the same virtual PR2.

5.1.3 Procedures. After providing informed consent and confirming that autoplayed videos worked on their computers, participants saw a video of the robot completing one of the warehouse tasks (Task X). The video began playing automatically, and a "proceed" button did not appear until the video ended. After responding to an attention check question, they were asked to evaluate a *different* warehouse task (Task Y) which they did not see performed in a video. Participants answered questions about capability, trust, and task assignment. This pattern was repeated with four total different X & Y task pairs. After answering these questions for all four pairs, participants rated how similar they found tasks X & Y to be for every pair of tasks. The task pairs that were analyzed are presented in Table 1. Participants answered a final attention check question, a short demographics questionnaire, and then were compensated.

Watched video of	Answered trust questions about
Welcome delivery people to the warehouse	Communicate product information with delivery people
Fix broken equipment	Close boxes with a drill
Identify broken products	Stack boxes
Place products in boxes	Label items

Table 1: Pairs of tasks that participants saw in Study 3.

5.1.4 Measures. The trust and capability questions were the same as the ones used in Studies 1 and 2. For the task similarity ratings, participants saw one of the 12 task pairing combinations at a time and rated their similarity on a 6-point "Very dissimilar" to "Very similar" Likert scale.

5.2 Results

To check if the level of similarity between a task that the participant had seen the robot complete and a task that they had not seen the robot complete affected perceptions of capability and trust for the unseen task, we ran linear regressions. Results indicated that there was no overall effect of similarity ratings on either capability perceptions ($F(1,206) = 1.495, p = .2228$, adjusted $R^2 = .0024$) or trust ($F(1,206) = 2.103, p = .1485$, adjusted $R^2 = .0053$). There was no effect of either gender ($p = .0627$) or age ($p = .628$). Overall, the results support the environmental factors hypothesis. However, to check whether this lack of correlation was due to individual differences in participants, we made scatterplots for each participant of their trust and similarity scores. Of our 52 participants, 12 had a positive correlation between task similarity and trust, 23 had a negative correlation, and 17 had no correlation (see Fig. 4 for representative plots for each of these types of correlations).

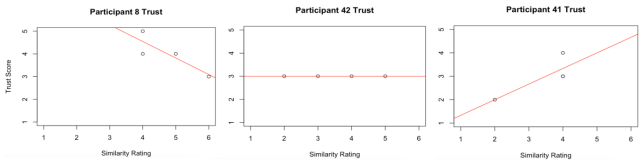

Figure 4: Representative examples of participants who had a negative correlation (Participant 8), no correlation (Participant 42), and a positive correlation (Participant 41) between task similarity and trust.

5.3 Discussion

While there was overall no correlation between task similarity ratings and trust scores, the lack of an effect was the result of significant individual differences, rather than being representative of most of the participants' scores. The wide spread of individual differences indicates that there is no common factor that explains how individuals make trust (and capability) judgments across tasks. For people with a positive correlation, the task factors hypothesis—that transfer happens across tasks that are similar—would explain the results. For those with no correlation, the environmental factors hypothesis—that transfer happens evenly across all tasks done in the same environment—would be accurate. It is not clear what kind of reasoning people with a negative correlation may be employing (e.g., perhaps they believe robots have that only do similar tasks have a narrow skill set and are generally less capable).

6 GENERAL DISCUSSION

In summary, we were initially motivated by understanding "what assumptions about a robot's capability and trust in the robot transfer from one environment to another after having seen the robot behave and perform tasks in the first environment". We hypothesized that participants who interacted with the robot in one environment would perceive the robot as more capable and trust it more in a second environment; they would remember what they observed in the first environment, and would apply that knowledge to assessments in the second. However, this did not turn out to be the case.

We proposed two alternative explanations: that the robot performing well in the first environment aligned with people's default models that robots are generally capable and work well, and therefore information further confirming this did not change baseline assumptions; or, alternatively, that the default model of the robot working well outweighs actual experience with the robot. We tested these hypotheses by adding a third experimental condition in which the robot's behavior was faulty in the first task. We predicted that, in line with the first explanation, this would go against people's default models of robots working well, and would result in lowered assessments of capability and trust in similar tasks as compared to the Control condition. However, the results did support this hypothesis, as participants in the Faulty condition rated the robot the same nearly across the board as those in the Functioning and Control condition.

We then ran a third study to disentangle whether task factors or environmental factors were more likely to affect trust transfer across tasks. We found no significant effects due to significant individual differences, with some participants' ratings in line with Soh et al.'s findings that increased task similarity led to increased trust [28–30]. For others, the fact that the robot is situated in a particular environment may lend credence to its capability at all tasks in the environment, as is perhaps supported by Robinette et al.'s work [22].

These individual differences indicate that there may not be a single factor (or even small set of factors) that robot designers can utilize to prompt all interactants to make the appropriate task transfer inferences and trust robots with new tasks in new environments. This is a critical insight for future HRI research, for it is highly desirable for many HRI contexts that people be able to make the right kinds of assumptions about what tasks robot may be able to perform, and how well, in different situations based on what they know about the robot from their experience with it For example, think of firefighters employing robots in search and rescue missions in new environments, or even household robots that are taken to different apartments with different kitchenware, appliances, etc.

In a next step, it would then be important to determine additional individual human factors such as the subjects' knowledge about robots, their past interactions with different types of robots, their overall dispositions towards robots (e.g., where and when they thought robots should be employed), and others. Those factors would then serve as co-variates in the above and similar studies, ideally helping to find systematic correlations with the different trust rating trends. And it would be important to then also vary the robot's appearance to exclude the possibility that the PR2 robot employed in all of our studies is somehow peculiar and that the same effects would not have been obtained with different robot types. Finally, while VR-based interaction studies are immersive and allow for HRI in otherwise unattainable environments (or with unattainable robots), it would be important to confirm that the employed experimental paradigm also holds true with physical robots co-located with subjects in physical environments.

7 CONCLUSION

We set out to explore how people's experiences with a robot in one environment affects their perceptions of it in a different environment; specifically, we were interested in whether people would transfer any assumptions they had made about the robot's capability and their trust in it from the first environment into the second environment. In our first study, we found no difference in people's perceptions of capability and trust in the robot whether they had prior experience with the robot in one environment before moving on to a second environment or not. We then hypothesized that this was because a robot working well in the first environment fits people's default model of robots *generally* working well; however, this hypothesis was refuted in our second study where we found that interacting with a faulty robot still resulted in the same capability and trust ratings. We then tested two different hypotheses, that people's perceptions were either generally affected by task-related factors and transfer would happen across similar tasks, or environmental-related factors and transfer would happen evenly across all tasks in one environment. While we found overall support for the environmental hypothesis, we also found strong individual differences in task similarity and trust dynamics among our participants. Taken together, the three studies thus establish that trust transfer across tasks and environments is not uniform across subjects, but seems to depend essentially on individual factors that are currently unknown. Hence, the findings point to an urgent need to isolate what underlies the significant individual difference we observed, for making appropriate task transfer inferences across tasks and environment is a critical factor in many envisioned robotic applications.

ACKNOWLEDGMENTS

This work was in part funded by AFOSR grant #FA9550-18-1-0465.

REFERENCES

[1] Abdulaziz Abubshait and Eva Wiese. 2017. You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in psychology* 8 (2017), 1393.

[2] Rowel Atienza and Alexander Zelinsky. 2002. Active gaze tracking for human-robot interaction. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 261–266.

[3] Meia Chita-Tegmark, Janet M Ackerman, Matthias Scheutz, et al. 2019. Effects of assistive robot behavior on impressions of patient psychological attributes: Vignette-based human-robot interaction study. *Journal of medical Internet research* 21, 6 (2019), e13729.

[4] Anders BH Christensen, Christian R Dam, Corentin Rasle, Jacob E Bauer, Ramlo A Mohamed, and Lars Christian Jensen. 2019. Reducing overtrust in failing robotic systems. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 542–543.

[5] Daniel P Davison, Frances M Wijnen, Vicky Charisi, Jan van der Meij, Vanessa Evers, and Dennis Reidsma. 2020. Working with a social robot in school: a long-term real-world unsupervised deployment. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 63–72.

[6] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 73–80.

[7] Angeliki Dimitrokalli, George-Christopher Vosniakos, Dimitris Nathanael, and Elias Matsas. 2020. On the assessment of human-robot collaboration in mechanical product assembly by use of Virtual Reality. *Procedia Manufacturing* 51 (2020), 627–634.

[8] Chad Edwards, Autumn Edwards, Patric R Spence, and David Westerman. 2016. Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies* 67, 2 (2016), 227–238.

[9] Denise Y Geiskkovitch, Raquel Thiessen, James E Young, and Melanie R Glenwright. 2019. What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 48–56.

[10] Jennifer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* Ieee, 55–60.

[11] PA Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2020. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors* (2020), 0018720820922080.

[12] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[13] Kerstin Sophie Haring, David Silvera-Tawil, Katsumi Watanabe, and Mari Velonaki. 2016. The influence of robot appearance and interactive ability in HRI: a cross-cultural study. In *International conference on social robotics*. Springer, 392–401.

[14] Kerstin S Kerstin S Haring, Katsumi Watanabe, Mari Velonaki, Chad C Tossell, and Victor Finomore. 2018. FFAB—The form function attribution bias in human–robot interaction. *IEEE Transactions on Cognitive and Developmental Systems* 10, 4 (2018), 843–851.

[15] Matthias Kraus, Johannes Kraus, Martin Baumann, and Wolfgang Minker. 2018. Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

[16] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 463–464.

[17] Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2021. The Interplay Between Emotional Intelligence, Trust, and Gender in Human-Robot Interaction. *Int. J. Soc. Robotics* 13, 2 (2021), 297–309.

[18] Theresa Law, Josh de Leeuw, and John H Long. 2020. How Movements of a Non-Humanoid Robot Affect Emotional Perceptions and Trust. *International Journal of Social Robotics* (2020), 1–12.

[19] Theresa Law, Bertram F Malle, and Matthias Scheutz. 2021. A touching connection: how observing robotic touch can affect human trust in a robot. *International Journal of Social Robotics* 13, 8 (2021), 2003–2019.

[20] Theresa Law and Matthias Scheutz. 2021. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction* (2021), 27–57.

[21] Sau-lai Lee, Ivy Yee-man Lau, Sara Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2767–2772.

[22] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

[23] Alessandra Rossi, Patrick Holthaus, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2018. 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*. 246–252.

[24] Selma Sabanovic, Marek P Michalowski, and Reid Simmons. 2006. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control, 2006.* IEEE, 596–601.

[25] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[26] Tracy Sanders, Alexandra Kaplan, Ryan Koch, Michael Schwartz, and Peter A Hancock. 2019. The relationship between trust and use choice in human-robot interaction. *Human factors* 61, 4 (2019), 614–626.

[27] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–65.

[28] Harold Soh, Pan Shu, Min Chen, and David Hsu. 2018. The Transfer of Human Trust in Robot Capabilities across Tasks.. In *Robotics: Science and Systems*.

[29] Harold Soh, Pan Shu, Min Chen, and David Hsu. 2019. Trust Dynamics and Transfer across Human-Robot Interaction Tasks: Bayesian and Neural Computational Models.. In *IJCAI*. 6226–6230.

[30] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* 39, 2-3 (2020), 233–249.

[31] Patric R Spence, David Westerman, Chad Edwards, and Autumn Edwards. 2014. Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports* 31, 3 (2014), 272–280.

[32] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 acm/ieee international conference on human-robot interaction*. 263–264.

[33] Vincent Weistroffer, Alexis Paljic, Lucile Callebert, and Philippe Fuchs. 2013. A methodology to assess the acceptability of human-robot collaboration using virtual reality. In *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*. 39–48.

[34] Sangseok You and Lionel P Robert. 2018. Human–robot similarity and willingness to work with a robotic co-worker. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–260.