Human Trust in Robot Capabilities across Tasks

Pan Shu
National University of Singapore
panshu@comp.nus.edu.sg

Chen Min
National University of Singapore
chenmin@comp.nus.edu.sg

Indu Bodala
National University of Singapore
indu@comp.nus.edu.sg

Stefanos Nikolaidis
Carnegie Mellon University
snikolai@andrew.cmu.edu

David Hsu
National University of Singapore
dyhsu@comp.nus.edu.sg

Harold Soh
National University of Singapore
harold@comp.nus.edu.sg

ABSTRACT

In this work, we study how humans transfer or generalize trust in robot capabilities across tasks, even with limited observations. We present results from a human-subjects experiment using a real-world Fetch robot performing household tasks. In summary, we find that human trust generalization is influenced by perceived task similarity, difficulty, and robot performance.

1 INTRODUCTION

Human trust in automation plays a prominent role in shaping interaction, for example, in guiding an operator's allocation strategy [5]. Inappropriate levels of trust may lead to undesirable outcomes, both over-reliance and under-utilization [3].

In this late breaking report, we investigate the nature of human trust *generalization* or *transfer* across tasks of different categories and difficulties. Trust is a multidimensional concept, with many factors characterizing human trust in automation, e.g., the human's technical expertise and the complexity of the automation [4]. We adopt the definition of trust as a psychological attitude [1] and focus on *trust in robot capabilities*, i.e., the belief in a robot's ability or competence to complete a task. Capability is a primal factor in determining overall trust in robots and thus, the decision to rely on them [4]. Following earlier work in robotics (e.g., [6]), we assume that robots are not intentionally deceptive.

Our main findings support the intuition that human trust transfers across tasks, and similar tasks are more likely to share a similar level of trust. Moreover, observations of robot performance influence the dynamics of human trust not only in the observed task, but also similar tasks. Finally, trust transfer is asymmetric: trust transfers more easily to simpler tasks than to more difficult tasks.

These findings have important implications for human-robot collaboration. A recent study showed that incorporating a human trust model into robot decision-making improved team performance on a table clearing task [2]. To infer human trust accurately in a collaboration *across multiple tasks*, our results suggest that robots should leverage the similarity and difficulty of tasks.

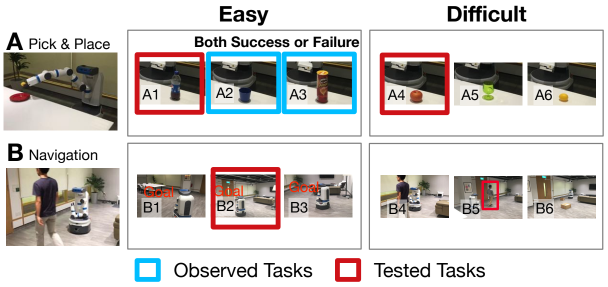


Figure 1: Trust Generalization Experiment Design. This work explores how human belief in robot capabilities generalizes across tasks given observations of robot performance on household tasks. Two categories of tasks were used: (A) picking and placing different objects, and (B) navigation in a room, potentially with people and obstacles. Participants were surveyed on their trust in the robot's ability to successfully perform three different tasks (red boxes) *before* and *after* being shown demonstrations of two tasks. The two observed tasks were always selected from the same cell (blue boxes; cell randomly assigned, with either both successes or both failures). The tested tasks were randomly selected from three different cells—the (i) same category and difficulty level, (ii) same category but different difficulty level, and (iii) different category but same difficulty level— compared to the observed tasks.

2 HUMAN-SUBJECTS EXPERIMENTS

Our general hypothesis is that human trust generalizes and evolves in a structured manner. More specifically, we hypothesize that:

- **H1:** Trust in the robot is more similar for tasks of the same category, compared to tasks in a different category.
- **H2:** Observations of robot performance have a greater affect on the *change in human trust* over similar tasks, i.e., in the same category, compared to dissimilar tasks.
- **H3:** Trust in a robot's ability to perform a task generalizes more readily to easier tasks, rather than to more difficult tasks.

2.1 Experimental Design

An overview of our experimental design is shown in Fig. 1. We explored three factors as independent variables: task category, task difficulty, and robot performance. Each independent variable consisted of two levels: two task categories, easy/difficult tasks, and robot success/failure.

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3177034

The robot used in this study was a real-world Fetch research robot with a 7-DOF arm. The tasks were typical household tasks, i.e., picking and placing different objects, and navigation in an indoor environment. We developed pre-programmed success and failure demonstrations of robot performance; for the navigation tasks, the robot was programmed to fail by moving to the wrong location. For picking and placing, the robot failed to grasp the target object.

The primary dependent variables were the participants' subjective trust in the robot r's capability to perform specific tasks. Participants indicated their trust given task x at time t, denoted as $T_{x,t}$,via a 7-point Likert scale in response to the agreement question: "*The robot is going to perform the task* $[x]$. *I trust that the robot can perform the task successfully*". From these task-dependent trust scores, we computed two derivative scores:

- **Trust Distance** $d_{T,t}(x,y) = |T_{x,t} - T_{y,t}|$, which is the 1-norm distance between scores for tasks x and y at time t.
- **Trust Change** $\Delta T_x(t_1,t_2) = |T_{x,t_1} - T_{x,t_2}|$, i.e., the 1-norm distance between the trust scores of task x at t_1 and t_2.

2.2 Study Procedure

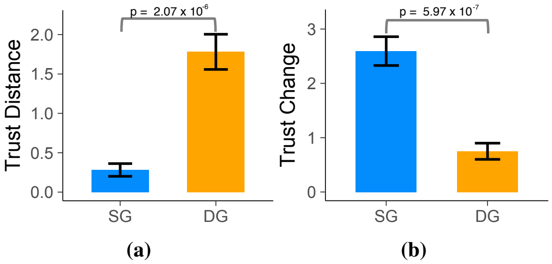
We recruited 32 individuals (Mean age: 24.09 years, $SD = 2.37$, 46.88% female) through an online advertisement and printed flyers on a university campus. After signing a consent form and providing standard demographic information, participants were introduced to the robot and followed through with the experiment's four stages:

(1) **Category and Difficulty Grouping:** To gain better control of the factors, participants were asked to group 12 tasks evenly into the four cells shown in Fig. 1. As such, chosen observations matched each participant's own prior estimations.
(2) **Pre-Observation Questionnaire:** Participants were asked to indicate their subjective trust on the three tested tasks using the measure instruments described above.
(3) **Observation of Robot Performance:** Participants were randomly assigned to observe two tasks from a specific category and difficulty, and were asked to indicate their trust if the robot were to repeat the observed task.
(4) **Post-Observation Questionnaire and Debrief:** Finally, participants were asked to re-indicate their subjective trust on the three tested tasks.

3 RESULTS

Our first set of results[1] are summarized in Fig. 2a, which show that tasks in the same category shared similar scores (supporting **H1**); their trust distances were significantly lower ($M = 0.28, SE = 0.081$) compared to tasks in other categories ($M = 1.78, SE = 0.22$); $t(31) = -5.82, p < 10^{-5}$.

Fig. 2b shows that the *change* in human trust due to performance observations of a given task was moderated by the perceived similarity of the tasks (**H2**). The trust change was significantly greater ($M = 2.59, SE = 0.26$) for tasks in the same category than otherwise ($M = 0.75, SE = 0.15$); $t(31) = 6.25$, $p < 10^{-6}$. Note also that the trust change for tasks in the different category was non-zero, $t(31) = 8.94, p < 10^{-6}$ for success and $t(31) = -8.35, p < 10^{-6}$ for failures respectively, indicating that trust generalizes even between task categories, albeit to a lesser extent.



(a) (b)

Figure 2: (a) Trust distance between a given task and tasks in the same category group (SG) compared to tasks in a different category (DG). Trust in robot capabilities was significantly more similar for tasks in the same group. (b) Trust change due to observations of robot performance. Trust increased (or decreased) significantly more for tasks in the SG versus DG.

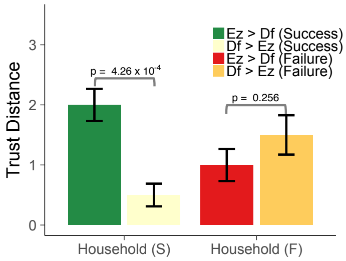


Figure 3: Trust distance between the observed task and a more difficult task (Ez → Df) against the converse, i.e., when generalizing to a simpler task (Df → Ez). Participants who observed successful demonstrations of a difficult task trusted the robot to perform simpler tasks, but not vice-versa.

Finally, we analyzed the relationship between perceived difficulty and trust generalization (**H3**) by first splitting the data into two conditions: participants who received successful demonstrations, and those that observed failures (Fig. 3). The trust distance was significantly less for tasks perceived to be easier than the observed task ($M = 2.0, SE = 0.27$), compared to tasks that were perceived to be more difficult ($M = 0.5, SE = 0.27$); $t(14) = 4.58, p < 10^{-3}$. For the failure condition, the results were not statistically significant, but suggest that the effect was reversed; belief in robot *inability* would transfer more to difficult tasks compared to simpler tasks.

REFERENCES

[1] Christiano Castelfranchi and Rino Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley Publishing, 1st edition, 2010.
[2] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. In *HRI'18*, page forthcoming. ACM, 2018.
[3] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.
[4] Bonnie M Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.
[5] Thomas B Sheridan and Robert T Hennessy. Research and modeling of supervisory control behavior. Technical report, NRC Committee on Human Factors, 1984.
[6] Anqi Xu and Gregory Dudek. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *HRI*, pages 221–228. ACM, 2015.

[1] Statistical significance for H1-H3 were assessed using both t-tests and non-parametric Wilcoxon signed-rank tests. Only t statistics are reported due to space constraints, but significance was achieved under both tests unless otherwise stated.