Multi-Task Trust Transfer for Human-Robot Interaction

Journal Title
XX(X):1–15

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Harold Soh, Yaqi Xie, Min Chen, and David Hsu

Abstract
Trust is essential in shaping human interactions with one another and with robots. This paper investigates how human trust in robot capabilities transfers across multiple tasks. It presents a human-subject study of two distinct task domains: a Fetch robot performing household tasks and a virtual reality simulation of an autonomous vehicle performing driving and parking maneuvers. The findings expand our understanding of trust and provide new differentiable models of trust evolution and transfer via latent task representations: a rational Bayes model, a data-driven neural network model, and a hybrid model that combines the two. Experiments show that the proposed models outperform prevailing models when predicting trust over unseen tasks and users. These results suggest that (i) task-dependent functional trust models capture human trust in robot capabilities more accurately, and (ii) trust transfer across tasks can be inferred to a good degree. The latter enables trust-mediated robot decision-making for fluent human-robot interaction in multi-task settings.

Keywords
Human-Robot Interaction, Trust, Human-Centered Robotics

Introduction

As robots enter our homes and workplaces, interactions between humans and robots become ubiquitous. *Trust* plays a prominent role in shaping these interactions and directly affects the degree of autonomy rendered to robots (Sheridan and Hennessy 1984). This has led to significant efforts in conceptualizing and measuring human trust in robots and automation (Muir 1994; Lee and Moray 1994; Castelfranchi and Falcone 2010; Yang et al. 2017).

A crucial gap, however, remains in understanding when and how human trust in robots *transfers* across multiple tasks based on the human's prior knowledge of robot task capabilities and past experiences. Understanding trust in the multi-task setting is crucial as robots transition from single-purpose machines in controlled environments—such as factory floors—to general-purpose partners performing diverse functions. The mathematical formalization of trust across tasks lays the foundation of trust-mediated robot decision making for fluent human-robot interaction. In particular, it leads to robot policies that mitigate under-trust or over-trust by humans when interacting with robots (Chen et al. 2018).

In this work, we take a first step on the question of formalizing trust transfer across tasks for human-robot interaction. We adopt the definition of trust as a psychological attitude (Castelfranchi and Falcone 2010) and focus on *trust in robot capabilities*, i.e., the belief in a robot's competence to complete a task. Capability is a primary factor in determining overall trust in robots (Muir 1994), and this work investigates how trust in robot capabilities varies and transfers across a range of tasks.

Our first contribution is a human-subject study ($n = 32$) where our goal is to uncover the role of task similarity and difficulty in the formation and dynamics of trust. We present results in two task domains: household tasks and autonomous driving (Fig. 1). The two disparate domains allow us to validate the robustness of our findings. We show that inter-task trust transfer depends on perceived task similarity, difficulty, and observed robot performance. These results are consistent across both domains, even though the robots and the contexts are markedly different: the household domain, involves a Fetch robot that navigates and that picks and places everyday objects, while the driving domain involves a virtual reality (VR) simulation of an autonomous vehicle performing driving and parking maneuvers. To our knowledge, this is the first work showing concrete evidence for trust transfer across tasks in the context of human-robot interaction. We have made our data and code freely available online for further research (Soh 2018).

Based on our experimental findings, we propose to conceptualize trust as a *context-dependent latent dynamic function*. This viewpoint is supported by prior socio-cognitive research showing the dependence of trust on task properties and on the agent to be trusted (Castelfranchi and Falcone 2010). We focus on characterizing the *structure* of this "trust function" and its *dynamics*, i.e., how it changes with observations of robot performance across tasks. An earlier version of this paper (Soh et al. 2018) presents two formal models: (i) a Bayesian Gaussian process (GP) (Rasmussen and Williams 2006) model, and (ii) a connectionist recurrent neural model based on recent

Department of Computer Science, School of Computing, National University of Singapore

Corresponding author:
Harold Soh, Dept of Computer Science, School of Computing, National University of Singapore
Email: harold@comp.nus.edu.sg

Figure 1. Experiment Task Domains: (Left) Household tasks with the Fetch Research Robot picking and placing objects (top left) and indoor navigation (bottom left) (Right) Autonomous Driving tasks in the Virtual Reality simulation system. Tasks included parking and various navigation scenarios.

advances in deep learning. The GP model explicitly encodes a specific assumption about how human trust evolves, via Bayes rule. In comparison, the neural model is data-driven and places few constraints on how trust evolves with observations. Both models leverage latent task space representations learned using word vector descriptions of tasks, e.g., "Pick and place a glass cup". Experiments show both models accurately predict trust across unseen tasks and users. This paper introduces a third model that combines the Bayesian and neural approaches and show that the hybrid model achieves improved predictions. All three models are differentiable and can be trained using standard off-the-shelf optimizers.

In comparison with prevailing computational models (e.g., Lee and Moray 1994; Xu and Dudek 2015), a key benefit of these trust models is their abilities to leverage inter-task structure in multi-task application settings. As *predictive* models, they can be operationalized in decision-theoretic frameworks to calibrate trust during collaboration with human teammates (Chen et al. 2018; Wang et al. 2016; Nikolaidis et al. 2017a; Huang et al. 2018). Trust calibration is crucial for preventing over-trust that engenders unwarranted reliance in robots (Robinette et al. 2016; Singh et al. 1993), and under-trust that can cause poor utilization (Lee and See 2004). To summarize, this paper makes the following key contributions:

- A novel formalization of trust as a latent dynamic function and efficient computational differentiable models that capture and predict human trust in robot capabilities across multiple tasks;
- Empirical findings from a human subjects study showing the influence of three factors on human trust transfer across tasks, i.e., perceived task similarity, difficulty, and robot performance;
- Systematic evaluation showing the proposed methods outperform existing methods, indicating the importance of modeling trust formation and transfer across tasks.

Background and Related work

Research into trust in robots (and automation) is a large interdisciplinary endeavor spanning multiple fields including human-factors, psychology, and human-robot interaction. This paper extends a prior conference version (Soh et al. 2018) with additional discussion and analyses of the human-subject experiment. In addition, we include a new computational trust model — that hybridizes the neural and Bayesian methods — with updated experimental results and expanded discussion on the computational models, learned task-spaces, and word-based task descriptions. In this section, we provide relevant background on trust and computational trust models.

Key Concepts and Definitions. Trust is a multidimensional concept, with many factors characterizing human trust in robots, e.g., the human's technical expertise and the complexity of the robot (Lee and Moray 1994; Muir 1994; Hancock et al. 2011). Of these factors, two of the most prominent are the performance and integrity of the machine. Similar to existing work in robotics (Xu and Dudek 2016, 2015), we assume that robots are not intentionally deceptive and focus on characterizing trust based on robot performance. We view trust as a belief in the competence and reliability of another agent to complete a task.

There are two types of trust that differ in their situation specificity. The first is dispositional trust or trust propensity, which is an individual difference for how willing one is to trust another. The second is situational or *learned* trust, that results from interaction between the agents concerned. For example, the more you use your new autonomous vehicle, the more you may learn to trust it. In this paper, we will be concerned mostly with situational trust in robots.

Trust Measurement. Trust is a latent dynamic entity, which presents challenges for measurement (Billings et al. 2012). Previous work has derived survey instruments and methods for quantifying trust, including binary measures (Hall 1996), continuous measures (Desai and Munjal 2012; Xu and Dudek 2016; Lee and Moray 1992), ordinal scales (Muir 1989; Hoffman et al. 2013; Jian et al. 2000) and an Area Under Trust Curve (AUTC) measure (Desai et al. 2013; Yang et al. 2017) which captures participant's trust through the entire interaction with the robot by integrating binary trust measures over time. In this paper, we use a self-reported measure of trust (similar to Xu and Dudek 2015) and Muir's questionnaire (Muir 1994).

Computational Models of Trust. Previous work has explored explanatory models (e.g., Castelfranchi and Falcone 2010; Lee and See 2004) and predictive models of trust. Recent models have focused on dynamic modeling, for example, a recent predictive model—OPTIMo (Xu and Dudek 2015)—is a Dynamic Bayesian Network with linear Gaussian trust updates trained on data gathered from an observational study. OPTIMo was shown to outperform an Auto-Regressive and Moving Average Value (ARMAV) model (Lee and Moray 1994), and stepwise regression (Xu and Dudek 2016). Because trust is treated as "global" real-valued scalar in these models, they are appropriate when tasks are fixed (or have little variation). However, as our results will show, trust can differ substantially between tasks. As such, we develop models that capture both the dynamic property of trust and its variation across tasks. We leverage upon recurrent neural networks that have been applied to a variety of sequential learning tasks (e.g., Soh et al. 2017) and

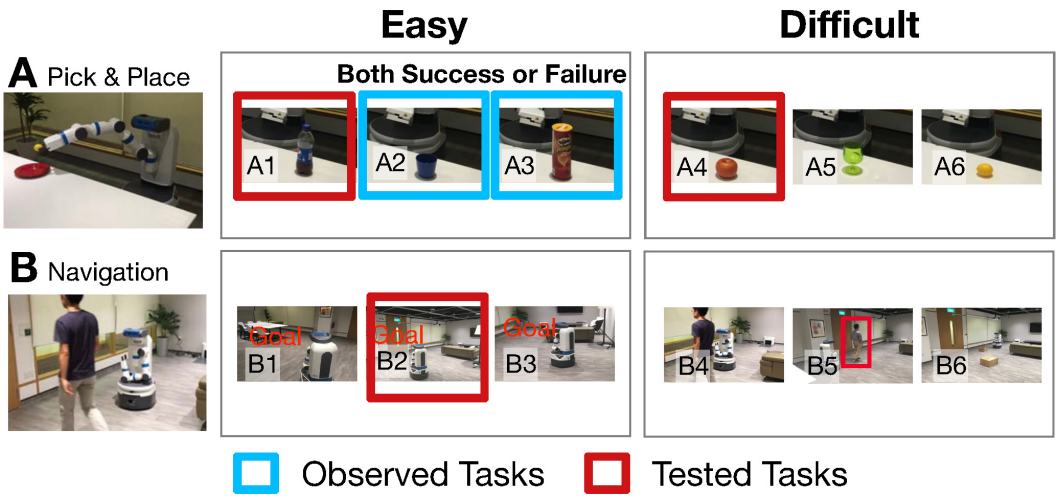


Figure 2. Trust Transfer Experiment Design. Two categories of tasks were used: (A) picking and placing different objects, and (B) navigation in a room, potentially with people and obstacles. Participants were surveyed on their trust in the robot's ability to successfully perform three different tasks (red boxes) *before* and *after* being shown demonstrations of two tasks. The two demonstrated/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.

online Gaussian processes that have been previously used in robotics (Soh and Demiris 2015, 2013, 2014).

Application of Trust Models. Trust emerges naturally in collaborative settings. In human-robot collaboration (Nikolaidis et al. 2017b,a), trust models can be used to enable more natural interactions. For example, Chen et al. (2018) proposed a decision-theoretic model that incorporates a predictive trust model, and showed that policies that took human trust into consideration led to better outcomes. The models presented in this work can be integrated into such frameworks to influence robot decision-making across different tasks.

Human Subjects Study

In this section, we describe our human subjects study, which was designed to evaluate if and when human trust transfers between tasks. Our general intuition was that human trust generalizes and evolves in a structured manner. We specifically hypothesized 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 compared to dissimilar tasks.
- **H3:** Trust in a robot's ability to perform a task transfers more readily to easier tasks, compared to more difficult tasks.
- **H4:** *Distrust* in the robot's ability to perform a task generalizes more readily to difficult tasks, compared to easier tasks.

Experimental Design

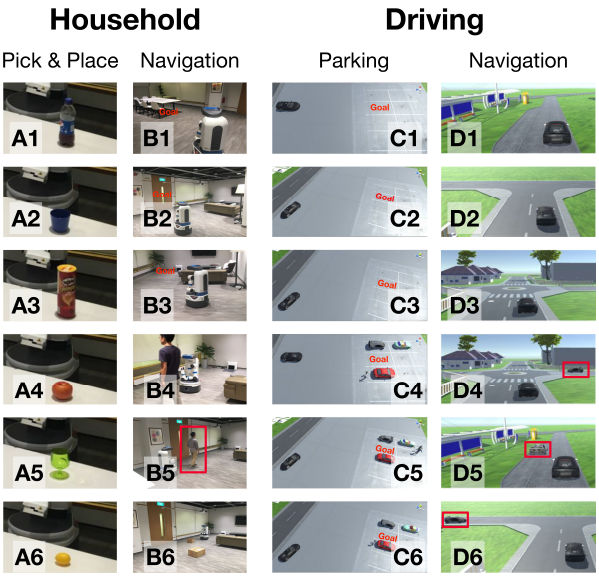
An overview of our experimental design is shown in Fig. 2. We explored three factors as independent variables: task category, task difficulty, and robot performance. Each independent variable consisted of two levels: two task categories, easy/difficult tasks, and robot success/failure. We used tasks in two domains, each with an appropriate robot (Fig. 3):

- **Household**, which included two common categories of household tasks, i.e., picking and placing objects, and navigation in an indoor environment. The robot used was a real-world Fetch research robot with a 7-DOF arm, which performed *live* demonstrations of the tasks in a lab environment that resembles a studio apartment.
- **Driving**, where we used a Virtual Reality (VR) environment to simulate an autonomous vehicle (AV) performing tasks such as lane merging and parking, potentially with dynamic and static obstacles. Participants interacted with the simulation system via an Oculus Rift headset, which provided a first-person viewpoint from the driver seat of the AV.

The robots were different in both settings and there were no cross-over tasks; in other words, the same experiment was conducted independently in each domain with the same protocol. Obtaining data from two separate experiments enabled us to discern if our hypotheses held in different contexts.

In both domains, we developed pre-programmed success and failure demonstrations of robot performance for all tasks. "Catastrophic" failures were avoided to mitigate complete distrust of the robot; for the household 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 autonomous car failed to park by stopping too far ahead of the lot, and failed to navigate (e.g., lane merge) by driving off the road and stopping (Fig. 4).

The primary dependent variables were the participants' subjective trust in the robot a's capability to perform specific



Domain	Category	ID	Task
Household	A. Pick & Place	1	a bottle of soda
		2	a plastic cup
		3	a can of chips
		4	an apple
		5	a glass cup
		6	a lemon
	B. Indoor Navigation	1	to the table
		2	to the door
		3	to living room
		4	with people moving around
		5	following a person
		6	while avoiding obstacles
Driving	C. Parking	1	Forwards, empty lot (aligned)
		2	Backwards, empty lot (misaligned)
		3	Forwards, empty lot (misaligned)
		4	Backwards, with cars (aligned)
		5	Backwards, with cars (misaligned)
		6	Forwards, with cars (misaligned)
	D. Navigation	1	Lane merge
		2	T-junction
		3	Roundabout
		4	Roundabout with other cars
		5	Lane merge with other cars
		6	T-junction with other cars

Figure 3. Tasks in the Household and Driving Domains. Tasks with IDs 1 to 3 are generally perceived to be easier than tasks labelled with IDs 4 to 6.

tasks. Participants indicated their degree of trust given robot a and task x at time t, denoted as $\tau^a_{x,t}$, 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*". In our form, the left-most point (1) indicated "*Strongly Disagree*" and the right-most point (7) indicated "*Strongly Agree*". From these task-dependent trust scores, we computed two derivative scores:

- **Trust distance across tasks** $d_{\tau,t}(x, x') = |\tau^a_{x,t} - \tau^a_{x',t}|$, i.e., the 1-norm distance between scores for x and x' at time t.
- **Trust change over time** $\Delta\tau^a_x(t_1, t_2) = |\tau^a_{x,t_1} - \tau^a_{x,t_2}|$, i.e., the 1-norm distance between the scores for x at t_1 and t_2.

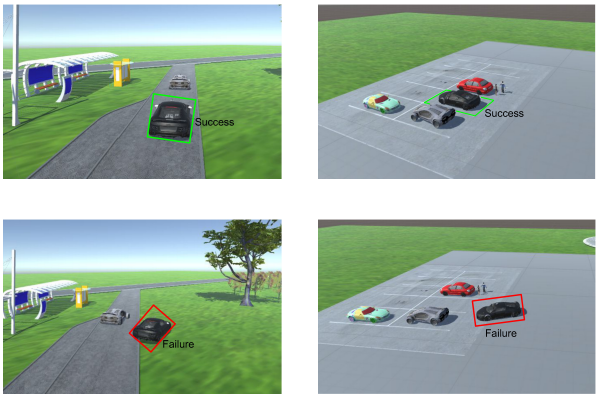


Figure 4. (Left Column) Autonomous car success (top) and failure (bottom) in the lane merge task. In the failure condition, the car drives off road and stops. (Right column) Autonomous car success (top) and failure (bottom) in the forwards parking task. In the failure condition, the car stops short of the parking spot.

As a general measure of trust, participants were also asked to complete Muir's questionnaire (Muir 1994; Muir and Moray 1996) pre-and-post exposure to the robot demonstrations. We also asked the participants to provide free-text justifications for their trust scores.

Robot Systems Setup

For both the Fetch Robot and Autonomous Driving simulator, we developed our experimental platforms using the Robot Operating System (ROS). On the Fetch robot, we used the MoveIt motion planning framework and the Open Motion Planning Library (Şucan et al. 2012) to pick and place objects, and the ROS Navigation stack for navigation in indoor environments.

The VR simulation platform was developed using the Unity 3D engine. Control of the autonomous vehicle was achieved using the hybrid A* search algorithm (Dolgov et al. 2010) and a proportional-integral-derivative (PID) controller.

Study Procedure

We recruited 32 individuals (Mean age: 24.09 years, $SD = 2.37$, 46% female) through an online advertisement and printed flyers on a university campus. Experiments were conducted in our lab where participants were shown live demonstrations of the Fetch robot performing the tasks, or observed the AV's behavior using the driving simulator. After signing a consent form and providing standard demographic data, participants were introduced to the robot. Specifically, they were provided information about the robot's parts and basic functions, and then asked questions to ensure that they were paying attention and understood the information. They then continued with the experiment's four stages:

1. **Category and Difficulty Grouping:** To gain better control of the factors, participants were asked to group the 12 tasks evenly into the four cells shown in Fig. 2. As such, chosen observations matched a participant's own prior estimations. We found that participant groupings were consistent — the same

grouping was observed across participants — but there was individual differences within each difficulty group, e.g., some participants thought picking and placing a lemon was comparatively more difficult than a glass cup.

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. The revised trust score is the baseline from which we evaluate the trust distance to tested tasks.
4. **Post-Observation Questionnaire and Debrief:** Finally, participants were asked to re-indicate their subjective trust on the three tested tasks, answered attention/consistency check questions*, and underwent a short de-briefing.

Results

In the following, we first report our primary findings using the task-dependent trust distance and change scores defined above. Then, we discuss the relationship between task-dependent trust and general trust (via comparison to the scores obtained from Muir's questionnaire). For the driving domain, one participant's results were removed due to a failure to pass attention/consistency check questions.

For the Household domain, an ANOVA showed that the effect of the category group on the trust distance was significant, $F(1,89) = 24.22, p < 10^{-5}$, as was the effect of difficulty, $F(1,89) = 14.05, p < 0.001$. The interaction between these two factors was moderately significant, $F(1,89) = 2.94, p = 0.089$. We also found a significant effect of category on trust change $F(1,89) = 24.89, p < 10^{-6}$, and of success/failure outcomes on trust change $F(1,89) = 10.04, p = 0.002$. Similar results were found for the driving domain.

Fig. 5 clearly shows that tasks in the same category (SG) shared similar scores (supporting **H1**); the post-observation trust distances (from the observed task to the tested tasks) are significantly lower $(M = 0.28, SE = 0.081)$ compared to tasks in other categories (DG) $(M = 1.78, SE = 0.22)$, $t(31) = -5.82$, $p < 10^{-5}$ for the household tasks. Similar statistically significant differences are observed for the driving domain, $t(30) = -2.755$, $p < 10^{-2}$. For both domains, we observe moderate effect sizes (≈ 1 on a Likert scale of 7), which suggests practical significance; the relative difference in trust may potentially affect subsequent decisions to delegate tasks to the robot.

Fig. 6 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 (between the pre-observation and post-observation trust scores for the three tested tasks) is significantly greater for tested tasks in the same group as the observed task (SG) than tasks in a different group (DG); $t(31) = 6.25, p < 10^{-6}$ for household and $t(30) = 3.46, p < 10^{-2}$ for driving.

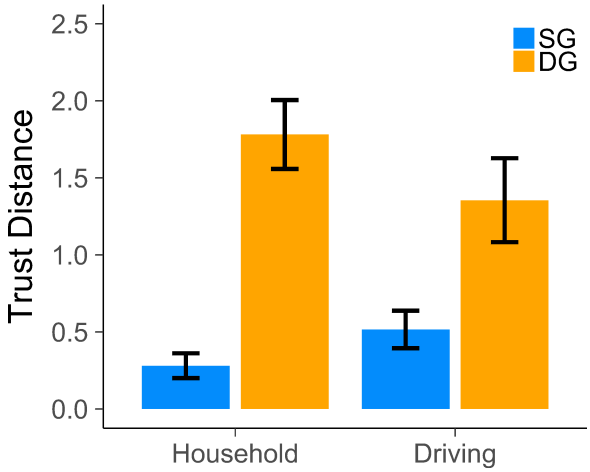


Figure 5. Trust distance between a given task and tasks in the same category group (SG) compared to tasks in a different category (DG). Trust in robot capabilities was significantly more similar for tasks in the same group.

Note also that the trust change for DG was non-zero (one-sample t-test, $p < 10^{-2}$ across both domains for successes and failures), indicating that trust transfers even between task categories, albeit to a lesser extent. Similar to the trust distances, the trust change effect sizes were moderately large indicating practical significance.

We analyzed the relationship between perceived difficulty and trust transfer (**H3**) by first splitting the data into two conditions: participants who received successful demonstrations, and those that observed failures (Fig. 7). For the success condition, the trust distance among the household tasks 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 $(M = 0.5, SE = 0.27)$, $t(14) = 4.58$, $p < 10^{-3}$. The hypothesis also holds in the driving domain, $M = 1.25$ $(SE = 0.25)$ v.s. $M = 2.43$ $(SE = 0.42)$, $t(14) = 3.6827, p < 10^{-3}$. For the failure condition (**H4**), the results were not statistically significant at the $\alpha = 1\%$ level, but suggest that the effect was reversed; belief in robot *inability* would transfer more to difficult tasks compared to simpler tasks.

Thus far, we have focussed on task-specific trust; a key question is how this task-dependent trust differs from a "general" notion of trust in the robot as measured by Muir's questionnaire. Fig. 8 sheds light on this question; overall, task-specific and general trust are positively correlated but the degree of correlation depends greatly on the similarity of the task to previous observations. In other words, while general trust is predictive of task-specific trust, it does not capture the range or variability of human trust across multiple tasks.

Summary of Findings and Discussion

Our main findings support the intuition that human trust transfers across tasks, but to different degrees. More specifically, similar tasks are more likely to share a similar

*Participants were asked several questions (e.g., what the last survey question was regarding) and to indicate their initial stated category/difficulty for a subset of the tasks.

Table 1. Effects of Participant Characteristics on Initial Trust and Trust Dynamics. Un-corrected p-values are shown. Bolded rows correspond to statistically significant coefficients at $\alpha = 0.1$ after Holm-correction.

	Initial Trust				Trust Change			
	coeff.	**std err**	**t-value**	**Pr($>|t|$)**	**coeff.**	**std err**	**t-value**	**Pr($>|t|$)**
Gender	0.43569	0.34425	1.266	0.211	-0.17572	0.10067	-1.745	0.086
Computer Usage								
Linear Term	-0.57557	0.31259	-1.841	0.071	0.20846	0.09142	2.280	0.026
Quadratic Term	0.46851	0.28246	1.659	0.103	0.06314	0.08261	0.764	0.448
Cubic Term	0.52029	0.27504	1.892	0.064	-0.07554	0.08043	-0.939	0.352
Experience with Robots								
Linear Term	**1.02462**	**0.35543**	**2.883**	**0.006**	-0.14597	0.10394	-1.404	0.166
Quadratic Term	0.06946	0.30585	0.227	0.821	-0.05689	0.08944	-0.636	0.527
Cubic Term	0.27462	0.22962	1.196	0.237	0.03136	0.06715	0.467	0.642
Experience with Video Games								
Linear Term	-0.54237	0.44622	-1.215	0.223	-0.23355	0.13049	-1.790	0.079
Quadratic Term	0.18745	0.32895	0.570	0.571	0.06359	0.09620	0.661	0.511
Cubic Term	-0.07300	0.26911	-0.271	0.787	-0.03880	0.07870	0.493	0.624
Quartic Term	-0.61713	0.23753	-2.598	0.012	**0.19424**	**0.06946**	**2.796**	**0.007**

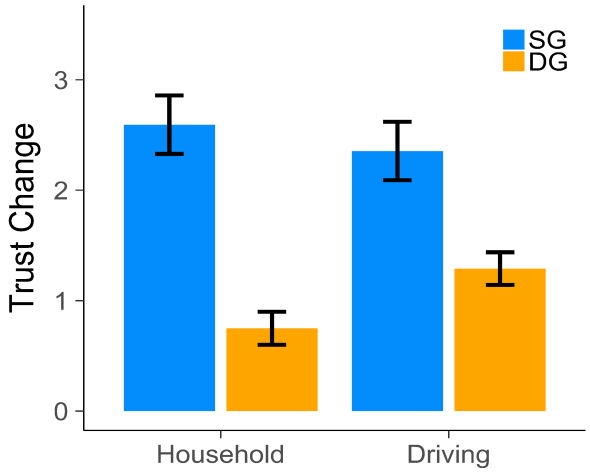


Figure 6. Trust change due to observations of robot performance. Trust increased (or decreased) significantly more for the tested tasks in the same group (SG) as the observed task versus tasks in different groups (DG).

level of trust (**H1**). Observations of robot performance changes trust both in the observed task, and also in similar yet unseen tasks (**H2**). Finally, trust transfer is asymmetric: positive trust transfers more easily to simpler tasks than to more difficult tasks (**H3**). These findings suggest that to infer human trust accurately in a collaboration across multiple tasks, robots should consider the similarity and difficulty of previous tasks.

Qualitative Analyses. Participant justifications for their trust scores were found to be consistent with the above findings. For example, a participant who was previously shown the robot successfully pick and place a plastic bottle and asked about her trust in the robot to pick and place a can, stated "*I trust this robot because the shape of the can of chips is similar to the bottle of soda*", whilst another participant who observed failures stated he distrusted the robot because the task was "*highly similar to the last two failed tasks*".

The justifications also revealed that some participants were more willing to trust initially (higher dispositional trust), e.g., "*Yes, I first gave the robot the benefit of the doubt on a task I saw that similar robot can perform some of the time; then I revised my trust completely based on what it actually did on a similar task*". Differences in perceived task difficulty also played a role in initial trust, "*I trust the robot because this seems like a simple enough task.*" and in trust transfer, for example, "*Robot failed much easier task of navigating around a stationary item, so I don't think it can follow a moving object*".

Trust and Participant Characteristics. Finally, we examine how participant characteristics may affect dispositional and situational trust. Specifically, we analyzed the effect of four independent variables—gender, amount of computer usage, prior experience with video games, and prior experience with robots—on the average initial trust and the average trust change. All four independent variables were self-reported; participants indicated their level of computer usage per week by selecting one of the following five choices: <10 hours, 10-20 hours, 20-30 hours, 30-40 hours, >40 hours. Prior experience with video games and robots was measured using the agreement questions, "*I'm experienced with video games*" and "*I'm experienced with robots*" using a 5-point Likert scale. We used the scores collected using Muir's questionnaire to compute the average initial trust and trust change. A polynomial contrast model (Saville and Wood 1991) was applied since the independent variables are ordinal and the true metric intervals between the levels are unknown[†]. We also ran tests against the task-specific trust but the results were not significant; this was potentially due to participants being exposed to very different tasks.

Table 1 summarizes our results. After correcting for family-wise error, we found a moderately significant association ($\alpha = 0.1$) between initial trust and prior experience with robots. Participants who had prior exposure to robots were more likely to trust the robots in our experiments. More experience with video games is significantly associated with trust changes. Although not statistically significant, it may also be negatively associated with initial trust (Holm-corrected $p = 0.12$). These results suggest that participant characteristics, such as their prior

[†]Polynomial contrast allows ordinal variables to enter not only linearly but also with higher order to better ascertain monotonic effects.

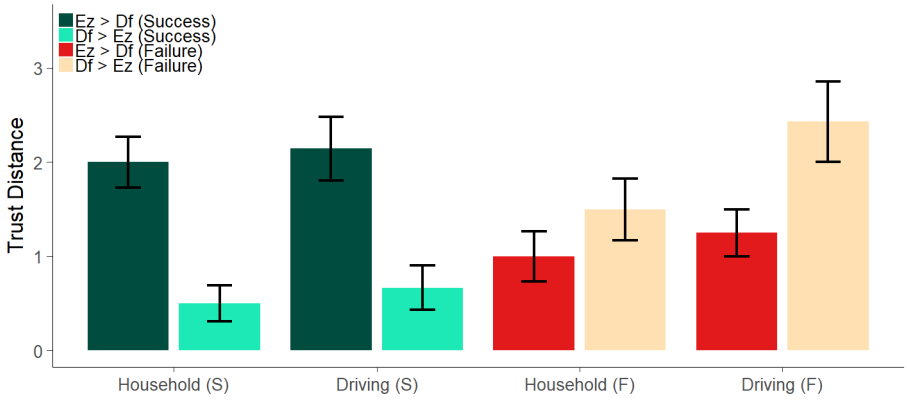


Figure 7. Trust distance between the observed task and a more difficult task (Ez → Df) against when generalizing to a simpler task (Df → Ez). Participants who observed successful demonstrations of a difficult task trusted the robot to perform simpler tasks, but not vice-versa.

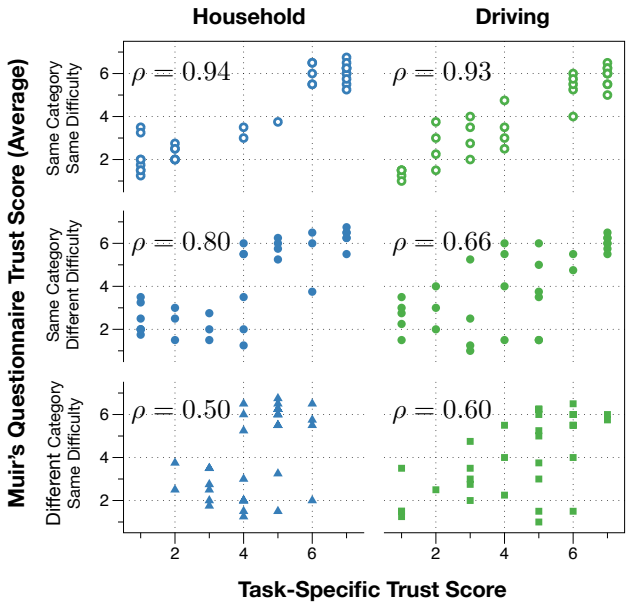


Figure 8. Muir (1994)'s Trust Score vs Task-specific trust scores (for tested tasks). The scores are positively correlated across the three task types, but with different strengths; the general measure is less predictive of task-specific trust for tasks in different categories (Pearson correlation, $\rho = 0.5 - 0.6$) compared to tasks with same category and difficulty ($\rho = 0.93 - 0.94$).

experience with technology, do play a role in trust formation and dynamics. However, the relationships do not appear straightforward and we leave further examination of these factors to future work.

Study Limitations. In this work, each participant only observed the robot performing two tasks; we plan to investigate longer interactions involving multiple trust updates in future work. Furthermore, our reported results are based on subjective self-assessments in non-critical tasks. We believe our results to remain valid when the robot's actions affect the human participant's goals. Our recent work (Xie et al. 2019; Chen et al. 2018) includes behavioral measures, such as operator take-overs and greater forms of risk (e.g., in the form of performance bonuses/penalties; these experiments also provide evidence for trust transfer across tasks..

Computational Models for Trust Across Multiple Tasks

The results from our human subjects study indicate that trust is relatively rich mental construct. Although we consider trust to be a useful information processing "bottleneck" in that it summarizes past experience with the robot, it does appear to be task-specific and hence, is more than a simple scalar quantity (as assumed in prior work Chen et al. 2018; Xu and Dudek 2016).

In this section, we present a richer model where trust is a *task-dependent latent dynamic function* $\tau_t^a(\mathbf{x}) : \mathbb{R}^d \rightarrow [0, 1]$ that maps task features, $\mathbf{x}$, to the continuous interval $[0, 1]$ indicating trustworthiness of the robot to perform the task. We assume that the task features are given and are sufficiently informative of the underlying tasks; for example, our experiments utilized word-vector features derived from English-language task descriptions, but visual features extracted from images or structured task descriptions may also be used.

This functional view of trust enables us to naturally capture trust differences across tasks, and can be extended to include other contexts; $\mathbf{x}$ can represent other factors, e.g., the current environment, robot characteristics, and observer properties. To model the dynamic nature of trust, we propose a Markovian function g that updates trust,

$$\tau_t^a = g(\tau_{t-1}^a, o_{t-1}^a) \tag{1}$$

where $o_{t-1}^a = (\mathbf{x}_{t-1}, c_{t-1}^a)$ is the observation of robot a performing a task with features $\mathbf{x}_{t-1}$ at time $t-1$ with performance outcome c_{t-1}^a. The function g serves to change trust given observations of robot performance, and as such, is a function over the space of trust functions. In this work, we consider binary outcomes $c_{t-1}^a \in \{+1, -1\}$ indicating success and failure respectively, but differences in performance can be directly accommodated via "soft labels" $c_{t-1}^a \in [-1, +1]$ without significant modifications to the presented methods.

The principle challenge is then to determine appropriate forms for τ_t^a and g. In this work, we propose and evaluate three different approaches: (i) a Bayesian approach where we model a probability distribution over latent functions via a Gaussian process, (ii) a connectionist approach utilizing a

recurrent neural network (RNN), and (iii) a hybrid approach that combines the aforementioned two methods.

Bayesian Gaussian Process Trust Model

In our first model, we view trust formation as a cognitive process, specifically, human function learning (Griffiths et al. 2009). We adopt a rational Bayesian framework, i.e., the human is learning about the robot capabilities by combining prior beliefs about the robot with evidence (observations of performance) via Bayes rule. More formally, let us denote the task at time t as $\mathbf{x}_t$, and the robot a's corresponding performance as c_t^a. Given binary success outcomes (where $c_t^a = 1$ indicates success), we introduce a latent function f^a and model trust in the robot as,

$$\tau_t^a(\mathbf{x}_t) = \int P(c_t^a = 1|f^a, \mathbf{x}_t) p_t(f^a) df^a \tag{2}$$

where $p_t(f^a)$ is the human's current belief over f^a, and $P(c_t^a|f^a, \mathbf{x}_t)$ is the likelihood of observing the robot performance c_t^a given the task $\mathbf{x}_t$. Intuitively, f^a can be thought of as a latent "unnormalized" trust function that has range over the real number line. Given an observation of robot performance, the human's trust is updated via Bayes rule,

$$p_t(f^a|\mathbf{x}_{t-1}, c_{t-1}^a) = \frac{P(c_{t-1}^a|f^a, \mathbf{x}_{t-1}) p_{t-1}(f^a)}{\int P(c_{t-1}^a|f^a, \mathbf{x}_{t-1}) p_{t-1}(f^a) df^a}, \tag{3}$$

where p_t is the posterior distribution over f^a.

To use this model, we need to specify the prior $p_0(f^a)$ and likelihood $P(c_i^a|f^a, \mathbf{x})$ functions. Similar to prior work in human function learning (Griffiths et al. 2009), we place a Gaussian process (GP) prior over f^a,

$$p_0(f^a) = \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{4}$$

where $m(\cdot)$ is the prior mean function, and $k(\cdot, \cdot)$ is the kernel or covariance function. The literature on GPs is large and we refer readers wanting more detail to Williams and Rasmussen (2006). In brief, a GP is a collection of random variables, of which any finite subset is jointly Gaussian. In this model, any given task feature $\mathbf{x}$ indexes a random variable representing the real function value $f^a(\mathbf{x})$ at specific location $\mathbf{x}$. The nice properties of Gaussians enable us to perform efficient marginalization, which makes the model especially attractive for predictive purposes. Note that the GP is completely parameterized by its mean and kernel functions, which we describe below.

Covariance Function. The kernel function is an essential ingredient for GPs and quantifies the similarities between inputs (tasks). Popular kernel functions include the squared exponential and Matérn kernels (Williams and Rasmussen 2006). Although our task features are generally high dimensional (e.g., the word features used in our experiments), we consider tasks to live on a low-dimensional manifold, i.e., a psychological task space. With this in mind, we use a projection kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp(-(\mathbf{x} - \mathbf{x}')^\top \mathbf{M} (\mathbf{x} - \mathbf{x}')) \tag{5}$$

with a low rank matrix $\mathbf{M} = \mathbf{\Lambda}\mathbf{L}\mathbf{\Lambda}^\top$ where $\mathbf{\Lambda} \in \mathbb{R}^{d \times k}$ and $\mathbf{L}$ is a diagonal matrix of length-scales capturing axis-aligned relevances in the projected task space.

Capturing Prior Estimations of Task Difficulty and Initial Bias. As our studies have shown, perceived difficulty results in an asymmetric transfer of trust (**H3**), which presents a difficulty for standard zero/constant-mean GPs given symmetric covariance functions. To address this issue, we explore two different approaches:

1. First, the mean function is a convenient way of incorporating a human's prior estimation of task difficulty; tasks which are presumed to be difficult (beyond the robot's capability) will have low values. Here, we have used a data-dependent linear function, $m(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$ where $\boldsymbol{\beta}$ is learned along with other GP parameters.
2. A second approach is to use pseudo-observations $\mathbf{x}^+$ and $\mathbf{x}^-$ and associated f^a's to bias the initial model. Intuitively, $\mathbf{x}^+$ ($\mathbf{x}^-$) summarizes the initial positive (negative) experiences that a person may have had. The pseudo-observations are implemented simply as pre-observed data-points that the models are seeded with, prior to any trust updates. Similar to $\boldsymbol{\beta}$, these parameters are learned during training. In our experiments, the pseudo-observations are trained using data from all the individuals in each training set, and thus, represent the "average" initial experience.

Both approaches allow the GP to accommodate the aforementioned asymmetry; the evidence has to counteract the prior mean function or pseudo-observations respectively.

Observation Likelihood. In standard regression tasks, the observed "outputs" are real-valued. However, participants in our experiments observed binary outcomes (the robot succeeded or failed) and thus, we use the probit likelihood (Neal 1997),

$$P(c_t^a|f^a, \mathbf{x}_t) = \Phi\left(\frac{c_t^a (f^a(\mathbf{x}_t) - m(\mathbf{x}_t))}{\sigma_n^2}\right) \tag{6}$$

where $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{t^2}{2}\right) dt$ is the CDF of the standard normal, and σ_n^2 is the noise variance. Here, $\Phi(y)$ is a response function that "squashes" the function value $y = f^a(\mathbf{x}) \in (-\infty, \infty)$ onto the range $[0, 1]$. Alternative likelihoods can be used without changing the overall framework.

Trust Updates via Approximate Bayesian Inference. Unfortunately, the Bayesian update (3) under the probit likelihood is intractable and yields a posterior process that is non-Gaussian. To address this problem and enable iterative trust updates, we employ approximate Bayesian inference: the posterior process is projected onto the closest GP as measured by the Kullback-Leibler divergence, $\mathrm{KL}(p_t||q)$, and q is our GP approximation (Csató and Opper 2002). Minimizing the KL divergence is equivalent to matching the first two moments of p_t and q, which can be performed analytically. The update equations in their natural parameterization forms are given by:

$$\mu_t(\mathbf{x}) = \boldsymbol{\alpha}_t^\top \mathbf{k}(\mathbf{x}) \tag{7}$$

$$k_t(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + \mathbf{k}(\mathbf{x})^\top \mathbf{C}_t \mathbf{k}(\mathbf{x}') \tag{8}$$

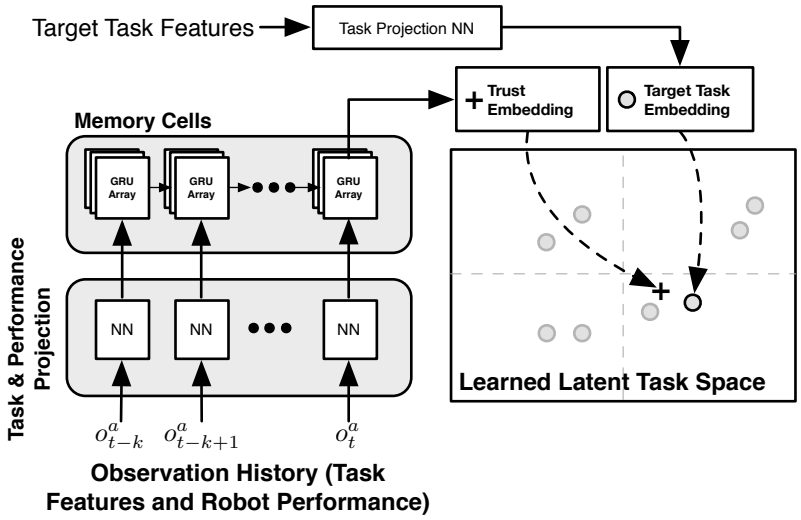


Figure 9. A High-level Schematic of the Neural Trust Model. The trust vector is updated using GRU cells as memory of previously observed robot performance. The model uses feed-forward neural networks to project tasks into a dot-product space in which trust for a task can be efficiently computed.

where $\boldsymbol{\alpha}$ vector and $\mathbf{C}$ are updated using:

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \gamma_1(\mathbf{C}_{t-1}\mathbf{k}_t + \mathbf{e}_t) \tag{9}$$

$$\mathbf{C}_t = \mathbf{C}_{t-1} + \gamma_2(\mathbf{C}_{t-1}\mathbf{k}_t + \mathbf{e}_t)(\mathbf{C}_{t-1}\mathbf{k}_t + \mathbf{e}_t)^\top \tag{10}$$

where $\mathbf{k}_t = [k(\mathbf{x}_1, \mathbf{x}_t), ..., k(\mathbf{x}_{t-1}, \mathbf{x}_t)]$, $\mathbf{e}_t$ is the t^{th} unit vector and the scalar coefficients b_1 and b_2 are given by:

$$\gamma_1 = \partial_{f^a} \log \int P(c^a_t | f^a, \mathbf{x}_t) df^a = \frac{c^a_i \partial\Phi}{\sigma_x \Phi} \tag{11}$$

$$\gamma_2 = \partial^2_{f^a} \log \int P(c^a_t | f^a, \mathbf{x}_t) df^a = \frac{1}{\sigma^2_x}\left[\frac{\partial^2\Phi}{\Phi} - \left(\frac{\partial\Phi}{\Phi}\right)\right] \tag{12}$$

where $\partial\Phi$ and $\partial^2\Phi$ are the first and second derivatives of Φ evaluated at $\frac{c^a_i(\mu_t(\mathbf{x}) - m(\mathbf{x}))}{\sigma_\mathbf{x}}$.

Trust Predictions. Given (9) and (10), predictions can be made with the probit likelihood (6) in closed-form:

$$\begin{aligned}\tau^a_t(\mathbf{x}) &= \int P(c^a = 1 | f^a, \mathbf{x}) p_t(f^a) df^a \\ &= \Phi\left(\frac{\mu_t(\mathbf{x}) - m(\mathbf{x})}{\sigma_\mathbf{x}}\right)\end{aligned} \tag{13}$$

where $\sigma_\mathbf{x} = \sqrt{\sigma^2_n + k_t(\mathbf{x}_i, \mathbf{x}_i)}$.

Neural Trust Model

The Gaussian process trust model is based on the assumption that human trust is essentially Bayesian in nature. However, this assumption may be too restrictive since humans are not thought to be fully rational or Bayesian[‡]. Here we consider an alternative "data-driven" approach based on recent advances in deep neural models.

The architecture of our neural trust model is illustrated in Fig. 9. We leverage a learned task representation or "embedding" space $Z \subseteq \mathbb{R}^k$ and model trust as a parameterized function over this space. The key idea is that (un-normalized) trust for a task is obtained via an inner-product between two components: a trust-vector $\boldsymbol{\theta}_t$ and a task representation $\mathbf{z}$. The trust vector $\boldsymbol{\theta}_t$ is a compressed representation of the human's prior interaction history (observations of robot performance o^a_t) and is derived using a recurrent neural network. The task representations $\mathbf{z}$ are derived using a learned function $f_z(\mathbf{x})$ over task features $\mathbf{x}$. To obtain a normalized trust score between $[0, 1]$, we use the sigmoid function,

$$\tau^a_t(\mathbf{x}; \boldsymbol{\theta}_t) = \text{sigm}(\boldsymbol{\theta}^\top_t f_z(\mathbf{x})) = \text{sigm}(\boldsymbol{\theta}^\top_t \mathbf{z}). \tag{14}$$

The trust function τ^a_t is fully parameterized by $\boldsymbol{\theta}_t$ and its linear form has benefits: it is efficient to compute given a task representation $\mathbf{z}$ and is interpretable in that the latent task space Z can be examined, similar to other dot-product spaces, e.g., word embeddings (Mikolov et al. 2013). Similar to the GP, Z can be seen as a psychological task space in which the similarities between tasks can be easily ascertained.

Task Projection. Whilst it is possible to train the model to learn separate task representations $\mathbf{z}$ for each task in the training set, this approach limits the model to only seen tasks. Our aim was to create a general model that potentially generalizes to new tasks. One could use the task features $\mathbf{x}$ directly in the trust function, but there is no guarantee that the task features would form a space in which dot products would give rise to meaningful trust scores. As such, we project observed task features $\mathbf{x}$ into Z via a nonlinear function, specifically, a fully-connected neural network,

$$\mathbf{z} = f_z(\mathbf{x}) = \text{NN}(\mathbf{x}; \theta_z) \tag{15}$$

where θ_z is the set of network parameters. Similarly, the robot's performance c^a is projected via a neural network, $\mathbf{c}^a = \text{NN}(c^a; \theta^a_c)$. During trust updates, both the task and performance representations are concatenated, $\hat{\mathbf{z}}_i = [\mathbf{z}; \mathbf{c}^a]$, as input to the RNN's memory cells.

[‡]Moreover, whether brains are truly Bayesian remains a matter of debate within the cognitive sciences (Bowers and Davis 2012).

Trust Updating via Memory Cells. We model the trust update function g using a RNN with parameters θ_g,

$$\boldsymbol{\theta}_t = \text{RNN}(\boldsymbol{\theta}_{t-1}, \hat{\mathbf{z}}_{t-1}; \theta_g). \tag{16}$$

In this work, we leverage on the Gated Recurrent Unit (GRU) (Cho et al. 2014), which is a variant of long short-term memory (Hochreiter and Schmidhuber 1997) with strong empirical performance (Jozefowicz et al. 2015). In brief, the GRU learns to control two internal "gates"—the update and reset gates—that affect what it remembers and forgets. Intuitively, the previous hidden state is forgotten when the reset gate's value nears zero. As such, cells that have active reset gates have learnt to model short-term dependencies. In contrast, cells that have active update gates model long-term dependencies (Cho et al. 2014). Our model uses an array of GRU cells that embed the interaction history up to time t as a "memory state" $\mathbf{h}_t$, which serves as our trust parameters $\boldsymbol{\theta}_t$.

More formally, a GRU cell k that has state $h_{t-1}^{(k)}$ and receives a new input $\hat{\mathbf{z}}_t$, is updated via

$$h_t^{(k)} = (1 - v_t^{(k)}) h_{t-1}^{(k)} + v_t^{(k)} \tilde{h}_t^{(k)}, \tag{17}$$

i.e., an interpolation of its previous state and a candidate activation $\tilde{h}_t^{(k)}$. This interpolation is affected by the update gate $v_t^{(k)}$, which is parameterized by matrices $\mathbf{W}_v$ and $\mathbf{U}_v$,

$$v_t^{(k)} = \text{sigm}\big([\mathbf{W}_v \hat{\mathbf{z}}_t + \mathbf{U}_v \mathbf{h}_{t-1})]_k\big). \tag{18}$$

The candidate activation $\tilde{h}_t^{(k)}$ is given by

$$\tilde{h}_t^{(k)} = \tanh([\mathbf{W}\hat{\mathbf{z}}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})]_k) \tag{19}$$

where $\odot$ denotes element-wise multiplication. The reset gate $r_t^{(j)} = [\mathbf{r}_t]_k$ is parameterized by two matrices $\mathbf{W}_r$ and $\mathbf{U}_r$,

$$r_t^{(k)} = \text{sigm}\big([\mathbf{W}_r \hat{\mathbf{z}}_t + \mathbf{U}_r \mathbf{h}_{t-1})]_k\big) \tag{20}$$

A GP-Neural Trust Model

Both the neural and Bayesian models assume Markovian trust updates and that trust summarizes past experience with the robot. They differ principally in terms of the inherent flexibility of the trust updates. In the RNN model, the update parameters, i.e., the gate matrices, are learnt. As such, it is able to adapt trust updates to best fit the observed data. However, the resulting update equations do not lend themselves easily to interpretation. On the other hand, the GP employs fixed-form updates that are constrained by Bayes rule. While this can hamper predictive performance (humans are not thought to be fully Bayesian), the update is interpretable and may be more robust with limited data.

A natural question is whether we can formulate a "structured" trust update that combines the simplicity of the Bayes update, while allowing for some additional flexibility. Here, we examine a variant of the GP model that incorporates a neural component in the mean function update. In particular, we modify Eq. (9) with an additional term:

$$\begin{aligned}\boldsymbol{\alpha}_t =& \boldsymbol{\alpha}_{t-1} + \gamma_1(\mathbf{C}_{t-1}\mathbf{k}_t + \mathbf{e}_t) + \\ & u(\boldsymbol{\alpha}_{t-1}, \mathbf{C}_{t-1}\mathbf{k}_t, \boldsymbol{\Lambda}\mathbf{x}_{t-1}, c_{t-1}^a)\end{aligned} \tag{21}$$

where $u(\cdot)$ models any residual not fully captured by the Bayes update. The function u takes as input the previous mean parameters $\boldsymbol{\alpha}_{t-1}$, $\mathbf{C}_{t-1}\mathbf{k}_t$, the latent task vector $\mathbf{z}_{t-1} = \boldsymbol{\Lambda}\mathbf{x}_{t-1}$, and the robot performance c_{t-1}^a. The key idea here is that a neural component may modify the posterior distribution in a data-driven (non-Bayesian) manner to better capture the intricacies of human trust updates. In our experiments, $u(\cdot)$ is a simple feed-forward neural network, but alternative models can be used without changing the overall framework. For example, using a GRU-based network would enable non-Markovian updates and may further improve performance.

Experiments

Our experiments were designed to establish if the proposed trust models that incorporate inter-task structure outperform existing baseline methods. In particular, we sought to answer three questions:

- **Q1** Is it necessary to model trust transfer, i.e., do the proposed function-based models perform better than existing approaches when tested on unseen participants?
- **Q2** Do the models generalize to unseen tasks?
- **Q3** Is it necessary to model differences in initial bias, specifically perceptions of task difficulty?
- **Q4** Does incorporating additional flexibility into the GP trust updates improve performance?

Experimental Setup

To answer these questions, we conducted two separate experiments. **Experiment E1** was a variant of the standard 10-fold cross-validation where we held-out data from 10% of the participants (3 people) as a test set. This allowed us to test each model's ability to generalize to unseen participants on the same tasks. To answer question **Q2**, we performed a leave-one-out test on the tasks (**Experiment E2**); we held-out all trust data associated with one task and trained on the remaining data. This process yielded 12 folds, one per task.

Trust Models. We evaluated six models in our experiments:

- **GP**: A constant-mean Gaussian process trust model;
- **PMGP**: The GP trust model with prior mean function;
- **POGP**: The GP trust model with prior pseudo-observations;
- **RNN**: The neural RNN trust model;
- **GPNN**: The Bayesian GP-neural trust model with prior pseudo-observations;
- **LG**: A linear Gaussian trust model similar to the updates used in OPTIMo (Xu and Dudek 2015);
- **CT**: A baseline model with constant trust.

The baseline CT and LG models did not utilize task features as they do not explicitly consider trust variation across tasks. The general LG model applies linear Gaussian updates:

$$p(\tau_t^a | \tau_{t-1}^a, c_{t-1}^a, c_{t-2}^a) = \mathcal{N}\left(\mathbf{w}_{\text{LG}}^\top \begin{bmatrix} \tau_{t-1}^a \\ c_{t-1}^a \\ c_{t-1}^a - c_{t-2}^a \end{bmatrix}, \sigma_{\text{LG}}^2\right) \tag{22}$$

where $\mathbf{w}_{\text{LG}}$ and σ^2_{LG} are learned parameters. In our dataset, the robot a's performance in the two time steps was the same (both success/failure). Hence, the updated trust only depends on the previous trust τ^a_{t-1} and robot performance c^a_{t-1}.

We implemented all the models using the PyTorch framework (Paszke et al. 2017). Preliminary cross-validation runs were conducted to find good parameters for the models. The RNN used a two layer fully-connected neural network with 15 hidden neurons and Tanh activation units to project tasks to a 30-dimensional latent task space (Eqn. (15)). The trust updates, Eqn. (16), were performed using two layers of GRU cells. A smaller 3-dimensional latent task space was used for the GP models. GP parameters were optimized during learning, except the length-scales matrix, which was set to the identity matrix $\mathbf{L} = \mathbf{I}$; fixing $\mathbf{L}$ resulted in a smoother optimization process. For the GPNN, $u(\cdot)$ was set as a simple two-layer feed-forward neural network with 20 neurons-per-layer and Tanh activation units.

Datasets. The models were trained using the data collected in our human subjects study. The RNN and GP-based models were *not* given direct information about the difficulty and group of the tasks since this information is typically not known at test time. Instead, each task was associated with a 50-dimensional GloVe word vector (Pennington et al. 2014) computed from the task descriptions in Fig. 3 (the average of all the word vectors in each description). Complete task descriptions and code to derive the features are available in the online supplementary material (Soh 2018).

Training. In these experiments, we predict how each individual's trust is dynamically updated. The tests are not conducted with a single "monolithic" trust model across all participants. Rather, training entails learning the latent task space and model parameters, which are shared among participants, e.g., $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ for the PMGP and the gate matrices for the GRU. However, each participant's model is updated *only* with the tasks and outcomes that the participant observes.

To learn these parameters, all models are trained "end-to-end". We applied maximum likelihood estimation (MLE) and optimized model parameters $\boldsymbol{\theta}$ using the Bernoulli likelihood of observing the normalized trust scores (as soft labels):

$$\mathcal{L}(\theta) = -\sum \hat{\tau}^a \log(1-\tau^a(\mathbf{x}))+ \tag{23}$$

$$(1-\hat{\tau}^a)\log(1-\tau^a(\mathbf{x})) \tag{24}$$

where $\hat{\tau}^a$ is the observed normalized trust score. In more general settings where trust is *not* observed, the models can be trained using observed human actions, e.g., in (Chen et al. 2017). We employed the ADAM algorithm (Kingma and Ba 2014) for a maximum of 500 epochs, with early stopping using a validation set comprising 15% of the training data.

Evaluation Evaluation is carried out on both pre-and-post-update trust. For both experiments, we computed two measures: the average Negative Log-likelihood (NLL) and Mean Absolute Error (MAE). However, we observed that these scores varied significantly across the folds (each participant/task split). To mitigate this effect, we also computed relative Difference from Best (DfB) scores: $\text{NLL}_{\text{DfB}}(i,k) = \text{NLL}(i,k) - \text{NLL}^*(i)$, where $\text{NLL}(i,k)$ is the NLL achieved by model k on fold i and $\text{NLL}^*(i)$ is the best score among the tested models on fold i. MAE_{DfB} is similarly defined.

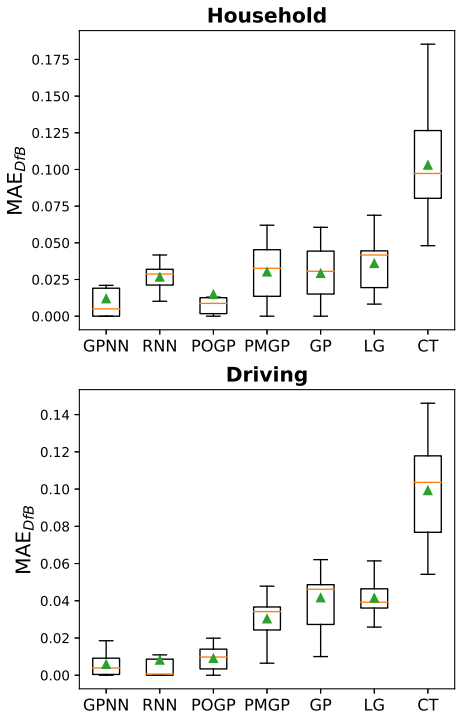


Figure 10. MAE_{DfB} scores for experiment **E1** with medians (lines) and means (triangles) shown. The proposed task-dependent trust models (GPNN, RNN, and POGP) models are superior at predicting trust scores on unseen test participants. The GPNN achieves the lowest average MAE_{DfB} scores across the two domains.

Results

Results for **E1** are summarized in Tbl. 2 with boxplots of MAE_{DfB} shown in Fig. 10. In brief, the GPNN, RNN, and POGP outperform the other models on both datasets across the performance measures. The POGP makes better predictions on the Household dataset, whilst the RNN performed better on the Driving dataset. The GPNN, however, obtains good performance across the datasets. In addition, the GP achieves better or comparable scores on average relative to LG and CT. Taken together, these results indicate that the answer to **Q1** is in the affirmative: accounting for trust transfer between tasks leads to better trust predictions.

Next, we turn our attention to **E2**, which is potentially the more challenging experiment. The GPNN, RNN, and POGP again outperform the other models (see Tbl. 3 and Fig. 11). Both models are able to make accurate trust predictions on *unseen* tasks (**Q2**), suggesting that (i) the word vectors are informative of the task, and (ii) the models learnt reasonable projections into the task embedding space.

To answer **Q3** (whether modeling initial human bias is required), we examined the differences between the POGP, PMGP, and GP. The PO/PMGP achieved lower or similar scores to the GP model across the experiments and domains, indicating that difficulty modeling enabled better performance. The pseudo-observation technique POGP

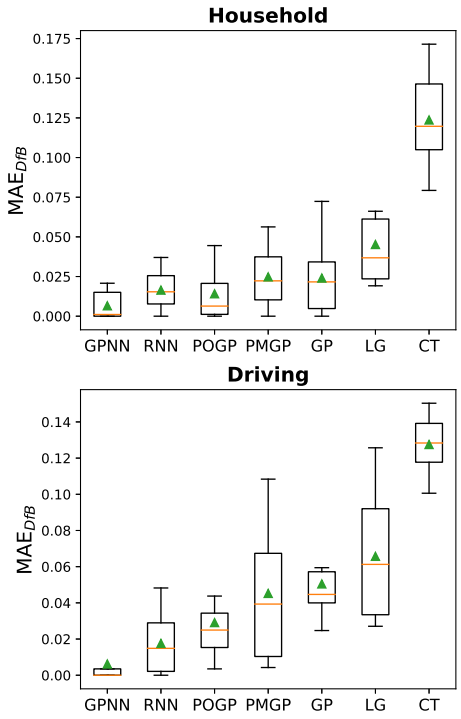


Figure 11. $\mathrm{MAE}_{\mathrm{DfB}}$ scores for experiment **E2** with medians (lines) and means (triangles) shown. The proposed task-dependent trust models (GPNN, RNN, and POGP) models are superior at predicting trust scores on unseen test tasks. Similar to Fig. 10, the GPNN achieves the lowest average $\mathrm{MAE}_{\mathrm{DfB}}$ scores across the two domains.

Table 2. Model Performance on Held-out Participants (Experiment **E1**). Average Negative log-likelihood (NLL) and Mean Absolute Error (MAE) scores shown with standard errors in brackets. Best scores in **bold**.

Models	Household		Driving	
	NLL	MAE	NLL	MAE
GPNN	**0.558** (**0.028**)	**0.158** (**0.011**)	0.555 (0.026)	**0.172** (**0.010**)
RNN	0.571 (0.023)	0.173 (0.010)	**0.549** (**0.024**)	0.175 (0.011)
POGP	0.558 (0.027)	0.161 (0.013)	0.553 (0.025)	0.176 (0.012)
PMGP	0.577 (0.019)	0.176 (0.010)	0.567 (0.018)	0.197 (0.011)
GP	0.575 (0.023)	0.175 (0.013)	0.588 (0.022)	0.208 (0.012)
LG	0.578 (0.023)	0.182 (0.011)	0.584 (0.022)	0.208 (0.011)
CT	0.662 (0.029)	0.249 (0.016)	0.649 (0.017)	0.266 (0.010)

always outperformed the linear mean function approach PMGP, suggesting initial bias is nonlinearly related to the task features. Potentially, using a non-linear mean function may allow PMGP to achieve similar performance to POGP.

Finally, we observed that the including additional flexibility in the GP mean update improved model performance (**Q4**). As stated above, the GPNN achieves similar or better performance on both datasets compared to the other GP variants.

Table 3. Model Performance on Held-out Tasks (Experiment **E2**). Average Negative log-likelihood (NLL) and Mean Absolute Error (MAE) scores shown with standard errors in brackets. Best scores in **bold**.

Models	Household		Driving	
	NLL	MAE	NLL	MAE
GPNN	**0.533** (**0.014**)	**0.156** (**0.007**)	0.542 (0.019)	**0.163** (**0.009**)
RNN	0.542 (0.012)	0.166 (0.006)	**0.531** (**0.016**)	0.174 (0.014)
POGP	0.542 (0.014)	0.164 (0.007)	0.562 (0.018)	0.186 (0.008)
PMGP	0.564 (0.014)	0.174 (0.009)	0.574 (0.015)	0.202 (0.008)
GP	0.551 (0.010)	0.174 (0.009)	0.586 (0.013)	0.207 (0.009)
LG	0.568 (0.014)	0.195 (0.009)	0.584 (0.013)	0.222 (0.010)
CT	0.669 (0.008)	0.273 (0.007)	0.661 (0.013)	0.284 (0.005)

Discussion

In summary, our experiments show that modeling trust correlations across tasks improves predictions. Our Bayesian and neural models achieve better performance than existing approaches that treat trust as a single scalar value. To be clear, neither model attempts to represent exact trust processes in the human brain; rather, they are computational analogues. Both modeling approaches offer conceptual frameworks for capturing the principles of trust formation and transfer. From one perspective, the GP model extends the single global trust variable used in Chen et al. (2018) and Xu and Dudek (2015) to a *collection* of latent trust variables. In the following, we highlight matters relating to the learned feature projections, changes in trust during a task, and the neural-GP trust updates.

Learned Trust-Task Spaces. Although the neural and Bayesian models differ conceptually and in details, they both leverage upon the idea of a vector task space $Z \subseteq \mathbb{R}^k$ in which similarities—and thus, trust transfer—between tasks can be easily computed. For the RNN, Z is a dot-product space. For the GP, similarities are computed via the kernel function; the kernel linearly projects the task features into a lower dimensional space ($\mathbf{z} = \mathbf{\Lambda x}$) where an anisotropic squared exponential kernel is applied. As an example, Figs. 12 and 13 show the learned GPNN latent task points for the Household and Driving domains respectively; we observe that tasks in the same group are clustered together. Furthermore, the easy and difficult tasks within each task group are also positioned closer together. This structure is consistent with the use of a squared exponential kernel where distances between the latent points determine covariance; the closer the points (more similar), the similar the latent function value at those points.

Generalization across Task Word Descriptions. In our experiments, we used word vector representations as task features, which we found to enable reasonable generalization across similar descriptions. For example, after seeing the

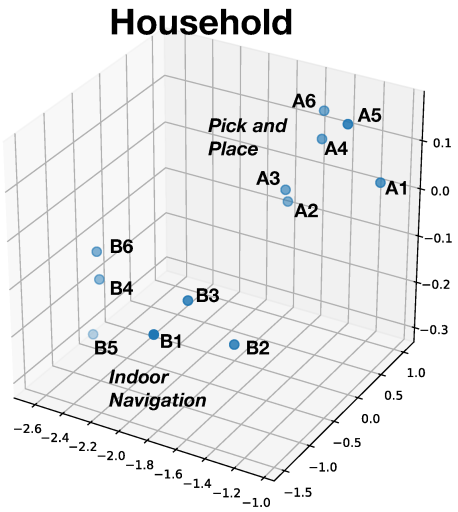


Figure 12. The task space for the Household domain where each point is a task. Tasks of a similar type and difficulty are clustered together; tasks labelled A correspond to Pick-and-Place tasks and B are Indoor Navigation tasks. The lower-numbered tasks (1-3) were considered easier by participants.

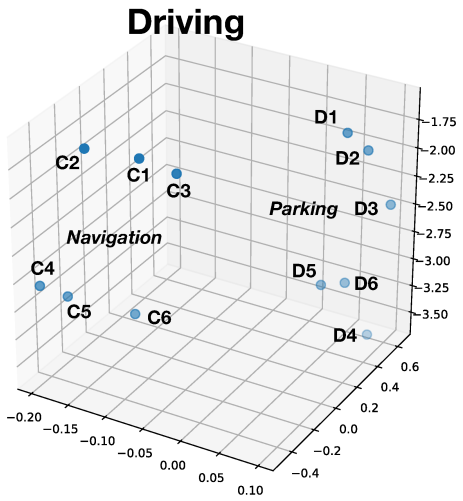


Figure 13. The task space for the Driving domain. Similar to Fig. 12 above, similar tasks are closer together. Points C corresponds to Navigation tasks (e.g., lane merging) and D are Parking tasks. The lower-numbered tasks (1-3) were considered easier by participants.

robot successfully navigate around obstacles, but *failing* to pick up a lemon, the model predicts sensible trust values for the following tasks:

- "*Navigate while following a path*": 0.81
- "*Go to the table*": 0.86
- "*Pick up a banana*": 0.61

We posit that this results from the fact that vector-based word representations are generally effective at capturing synonyms and word relationships. Given a latent space with sufficiently large dimensionality, we expect the model to scale to a larger number of task categories and domains; there is evidence that moderately-sized latent spaces ($<$ 1000) yield accurate models for complex tasks such as language translation (Sutskever et al. 2014) and image captioning (Ren et al. 2017). Given longer task descriptions, more sophisticated techniques from NLP (e.g., Bowman et al. 2016) beyond the simple averaging used in our experiments can be adapted to construct usable task features of reasonable length.

A prevailing issue is that the current word/sentence representations may not distinguish subtle semantics, e.g., the task features lack a notion of physical constraints. As such, the model may make unreasonable predictions when given task descriptions that are syntactically similar but semantically different. As an example, the same model predicts the human highly trusts the robot's capability to "*Navigate to the moon*" ($\tau^a = 0.83$). To remedy this issue, we can use alternative features; more informative vector-based features can be used without changing the methods described. Applying structured feature representations (e.g., graphs) would require different kernels and embedding techniques. Future work may also examine more sophisticated hierarchical space representations.

Trust Variations within a Task. The presented computational models are "event-based" in that trust was updated after each complete task performance. However, prior work has shown that trust can change even as the task is being carried out (Desai et al. 2013; Yang et al. 2017). To accommodate intra-task trust variability, the presented Bayesian and neural models can be easily altered to be updated at user-defined intervals with a corresponding observation. These modifications and studies to validate such models would make for interesting future work.

GP-Neural Updates. Finally, we sought to better understand the relationship between the Bayesian and neural components of the GPNN mean update. Was the neural network $u(\cdot)$ only making minor "corrections" to the trust update or was it playing a larger role? To answer this question, we compared the relative norms of the second term ($\eta_{\text{GP}} = \|\gamma_1(\mathbf{C}_{t-1}\mathbf{k}_t + \mathbf{e}_t)\|/\|\boldsymbol{\alpha}_{t-1}\|$) and third term ($\eta_{\text{NN}} = u(\boldsymbol{\alpha}_{t-1}, \mathbf{C}_{t-1}\mathbf{k}_t, \boldsymbol{\Lambda}\mathbf{x}_{t-1}, c^a_{t-1})/\|\boldsymbol{\alpha}_{t-1}\|$) on the RHS of Eq. (21) during updates across randomly sampled tasks. Fig. 14 shows a scatter plot of the relative norms of both components. We find a positive correlation (Kendall tau $= 0.2$, p-value $= 10^{-39}$), but the relationship is clearly nonlinear. Interestingly, η_{NN} could be relatively large when the η_{GP} was close to zero, indicating that neural component was playing a significant role in the trust update. We also experimented with completely removing the Bayesian portion of the update, but this modification had poorer performance, potentially due to limited data. This suggests that trust is not purely Bayesian and a non-trivial correction is needed to achieve better performance.

Conclusion

This paper takes a first step towards conceptualizing and formalizing predictive of human trust in robots across multiple tasks. It presents findings from a human-subjects study in two separate domains and shows the effects of task similarity and difficulty on trust formation and transfer.

The experimental findings leads to three novel models that capture the form and evolution of trust as a latent function. Our experiments show that the function-based models achieved better predictive performance on both unseen

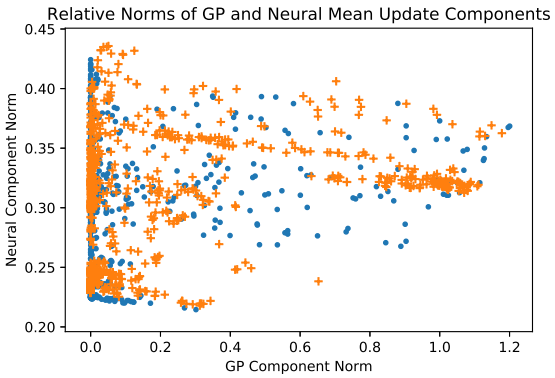


Figure 14. The relative norms of the GP component η_{GP} (x-axis) and neural component η_{NN} (y-axis) for updates across randomly sampled tasks. Blue points are for the Household domain and orange +'s for Driving. There is a general positive correlation between the norms, but the relationship is nonlinear.

participants and unseen tasks. These results indicate that (i) a task-dependent functional trust model more accurately reflects human trust across tasks, and (ii) it is possible to accurately predict trust transfer by leveraging upon a shared task space representation and update process.

Formalizing trust as a function opens up several avenues for future research. In particular, we aim to fully exploit this characterization by incorporating other contexts. Does trust transfer when the environment changes substantially or a new, but similar robot appears? Proper experimental design to elicit and measure trust is crucial. Our current experiments employ relatively short interactions with the robot and rely on subjective self-assessments. Future experiments could employ behavioral measures, such as operator take-overs and longer-term interactions where trust is likely to play a more significant role.

This work limits the investigation to trust resulting from complete observations of the robot's performance/capabilities. However, in real-world collaborative settings, the human user may not observe all the robot's successes or failures: how humans infer the robot's performance under conditions of partial observability remains an interesting open question. It is also essential to examine trust in the robot's "intention", e.g., its policy (Huang et al. 2018) and decision-making process. Arguably, trust is most crucial in new and uncertain situations whereby both robot capability and intention can influence outcomes. In very recent work (Xie et al. 2019), we have begun to examine how human mental models of these factors influence decisions to trust robots.

Finally, it is important to investigate how these trust models enhance human-robot interaction. In our current experiments, the human does not get involved in task completion. Our trust model can be used without modification in the collaborative setting where the human and the robot work together to complete a task, provided that the shared goal is unaffected by the change in trust. Embedding trust models in a decision theoretic framework enables a robot to adapt its behavior according to a human teammate's trust and as a result, promotes fluent long-term collaboration. We have begun a preliminary investigation using trust transfer models in a Partially-observable Markov Decision Process (POMDP), extending the work in (Chen et al. 2018). We are particularly interested in how trust models impacts decision-making in assistive tasks (Gombolay et al. 2018; Soh and Demiris 2015).

Acknowledgements

This work was supported in part by a NUS Office of the Deputy President (Research and Technology) Startup Grant and in part by MoE AcRF Tier 2 grant MOE2016-T2-2-068. Thank you to Indu Prasad for her help with the data analysis.

References

Billings DR, Schaefer KE, Chen JYC and Hancock PA (2012) Human-Robot Interaction: Developing Trust in Robots. *HRI '12* : 5–8.

Bowers JS and Davis CJ (2012) Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin* 138(3): 389.

Bowman SR, Vilnis L, Vinyals O, Dai AM, Jozefowicz R and Bengio S (2016) Generating sentences from a continuous space. *CoNLL 2016* : 10.

Castelfranchi C and Falcone R (2010) *Trust Theory: A Socio-Cognitive and Computational Model*. 1st edition. Wiley Publishing.

Chen M, Nikolaidis S, Soh H, Hsu D and Srinivasa S (2017) The role of trust in decision-making for human robot collaboration. In: *Workshop on Human-Centered Robotics, RSS*.

Chen M, Nikolaidis S, Soh H, Hsu D and Srinivasa S (2018) Planning with trust for human-robot collaboration. In: *HRI'18*. New York, NY, USA: ACM. ISBN 978-1-4503-4953-6, pp. 307–315.

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: *EMNLP*. pp. 1724–1734.

Csató L and Opper M (2002) Sparse on-line gaussian processes. *Neural Comp.* 14(3): 641–668.

Desai and Munjal (2012) Modeling trust to improve human-robot interaction .

Desai M, Kaniarasu P, Medvedev M, Steinfeld A and Yanco H (2013) Impact of robot failures and feedback on real-time trust. In: *HRI '13*. pp. 251–258.

Dolgov D, Thrun S, Montemerlo M and Diebel J (2010) Path planning for autonomous vehicles in unknown semi-structured environments. *IJRR* 29(5): 485–501.

Gombolay M, Yang XJ, Hayes B, Seo N, Liu Z, Wadhwania S, Yu T, Shah N, Golen T and Shah J (2018) Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research* 37(10): 1300–1316.

Griffiths TL, Lucas CG, Williams JJ and Kalish ML (2009) Modeling human function learning with Gaussian processes. *NIPS 21* : 553–560.

Hall R (1996) Trusting your assistant. In: *Proceedings of the 11th Knowledge-Based Software Engineering Conference*. IEEE Comput. Soc. Press. ISBN 0-8186-7680-9, pp. 42–51. DOI: 10.1109/KBSE.1996.552822.

Hancock PA, Billings DR, Schaefer KE, Chen JY, De Visser EJ and Parasuraman R (2011) A meta-analysis of factors affecting

trust in human-robot interaction. *Human Factors* 53(5): 517–527.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural computation* 9(8): 1735–80.

Hoffman RR, Johnson M and Bradshaw JM (2013) Trust in Automation. *IEEE Intelligent Systems* (13): 1541—-1672.

Huang SH, Bhatia K, Abbeel P and Dragan AD (2018) Establishing (appropriate) trust via critical states. In: *Workshop on Explainable Robotic Systems, HRI*.

Jian JY, Bisantz AM and Drury CG (2000) Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4(1): 53–71. DOI:10.1207/S15327566IJCE0401_04.

Jozefowicz R, Zaremba W and Sutskever I (2015) An empirical exploration of recurrent network architectures. In: *ICML*. pp. 2342–2350.

Kingma DP and Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Lee J and Moray N (1992) Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35(10): 1243–1270. DOI:10.1080/00140139208967392.

Lee JD and Moray N (1994) Trust, self-confidence, and operators' adaptation to automation. *Intl. J. of Human-Computer Studies* 40(1): 153–184.

Lee JD and See KA (2004) Trust in automation: Designing for appropriate reliance. *Human Factors* 46(1): 50–80.

Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J (2013) Distributed representations of words and phrases and their compositionality.

Muir BM (1989) *Operators trust in and percentage of time spent using the automatic controllers in a supervisory process control task*. University of Toronto. ISBN 0315510145.

Muir BM (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37(11): 1905–1922.

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

Neal RM (1997) Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026* .

Nikolaidis S, Nath S, Procaccia AD and Srinivasa S (2017a) Game-theoretic modeling of human adaptation in human-robot collaboration. In: *HRI'17*. New York, NY, USA: ACM. ISBN 978-1-4503-4336-7, pp. 323–331.

Nikolaidis S, Zhu YX, Hsu D and Srinivasa S (2017b) Human-robot mutual adaptation in shared autonomy : 294–302DOI: 10.1145/2909824.3020252. URL http://doi.acm.org/10.1145/2909824.3020252.

Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L and Lerer A (2017) Automatic differentiation in pytorch .

Pennington J, Socher R and Manning CD (2014) Glove: Global vectors for word representation. In: *EMNLP*. pp. 1532–1543.

Rasmussen CE and Williams CKI (2006) *Gaussian Processes for Machine Learning*. The MIT Press.

Ren Z, Wang X, Zhang N, Lv X and Li LJ (2017) Deep reinforcement learning-based image captioning with embedding reward. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 290–298.

Robinette P, Li W, Allen R, Howard AM and Wagner AR (2016) Overtrust of robots in emergency evacuation scenarios. In: *HRI'16*. pp. 101–108.

Saville DJ and Wood GR (1991) *Polynomial Contrasts*. New York, NY: Springer New York. ISBN 978-1-4612-0971-3, pp. 224–270.

Sheridan TB and Hennessy RT (1984) Research and modeling of supervisory control behavior. report of a workshop. Technical report, National Research Council Washington DC Committee on Human Factors.

Singh IL, Molloy R and Parasuraman R (1993) Automation-induced "complacency": Development of the complacency-potential rating scale. *The Intl. J. of Aviation Psychology* 3(2): 111–122.

Soh H (2018) Supplementary online material for predicting human trust in robot capabilities across tasks. https://github.com/crslab/human-trust-transfer.

Soh H and Demiris Y (2013) When and how to help: An iterative probabilistic model for learning assistance by demonstration. In: *IROS'13*. pp. 3230–3236.

Soh H and Demiris Y (2014) Spatio-temporal learning with the online finite and infinite echo-state gaussian processes. *IEEE Transactions on Neural Networks and Learning Systems* 26. DOI:10.1109/TNNLS.2014.2316291.

Soh H and Demiris Y (2015) Learning Assistance by Demonstration: Smart Mobility With Shared Control and Paired Haptic Controllers. *Journal of Human-Robot Interaction* 4(3): 76–100.

Soh H, Pan S, Min C and Hsu D (2018) The transfer of human trust in robot capabilities across tasks. In: *Proceedings of Robotics: Science and Systems*. Pittsburgh, Pennsylvania. DOI: 10.15607/RSS.2018.XIV.033.

Soh H, Sanner S, White M and Jamieson G (2017) Deep Sequential Recommendation for Personalized Adaptive User Interfaces. In: *IUI '17*. pp. 589–593. DOI:10.1145/3025171.3025207.

Şucan IA, Moll M and Kavraki LE (2012) The Open Motion Planning Library. *IEEE Robotics & Automation Magazine* 19(4): 72–82.

Sutskever I, Vinyals O and Le QV (2014) Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND and Weinberger KQ (eds.) *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 3104–3112.

Wang N, Pynadath DV and Hill SG (2016) Trust calibration within a human-robot team: Comparing automatically generated explanations. In: *HRI '16*. pp. 109–116.

Williams CK and Rasmussen CE (2006) *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA.

Xie Y, Prasad I, Ong D, Hsu D and Soh H (2019) Robot capability and intention in trust-based decisions across tasks. In: *HRI'19*.

Xu A and Dudek G (2015) Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In: *HRI*. ACM, pp. 221–228.

Xu A and Dudek G (2016) Towards Modeling Real-Time Trust in Asymmetric Human–Robot Collaborations. In: *Robotics Research*. pp. 113–129. DOI:10.1007/978-3-319-28872-7_7.

Yang XJ, Unhelkar VV, Li K and Shah JA (2017) Evaluating effects of user experience and system transparency on trust in

automation. In: *HRI*. pp. 408–416.