Trust Dynamics and Transfer across Human-Robot Interaction Tasks: Bayesian and Neural Computational Models

Harold Soh, Shu Pan, Min Chen and David Hsu
Dept. of Computer Science, National University of Singapore
{harold, panshu, chenmin, dyhsu}@comp.nus.edu.sg

Abstract
This work contributes both experimental findings and novel computational human-robot trust models for multi-task settings. We describe Bayesian non-parametric and neural models, and compare their performance on data collected from real-world human-subjects study. Our study spans two distinct task domains: household tasks performed by a Fetch robot, and a virtual reality driving simulation of an autonomous vehicle performing a variety of maneuvers. We find that human trust changes and transfers across tasks in a structured manner based on perceived task characteristics. Our results suggest that task-dependent functional trust models capture human trust in robot capabilities more accurately, and trust transfer across tasks can be inferred to a good degree. We believe these models are key for enabling trust-based robot decision-making for natural human-robot interaction.

1 Introduction
Trust forms the fabric of human-human relationships and, by extension, the relationships between humans and autonomous agents. For example, our decision to delegate tasks to another human or autonomous agent depends substantially on our trust in the agent [Xie et al., 2019]. As AI enters our daily life, human trust in autonomous agents will impact how these systems are used or misused. Trust calibration is crucial for preventing over-trust, which results in unwarranted reliance in robots [Robinette et al., 2016], or under-trust, which can cause poor utilization [Lee and See, 2004].

This extended abstract summarizes our recent work [Soh et al., 2018] that brings together empirical findings and new modeling tools to produce novel computational models of trust dynamics in human-robot interaction tasks. In contrast to prevailing approaches, e.g., [Lee and Moray, 1994; Xu and Dudek, 2015], our models leverage inter-task structure and are applicable across different tasks that a single agent must handle. As predictive models, they can be easily situated within a decision-theoretic framework, such as the partially observable Markov decision process [Kaelbling et al., 1998], to calibrate trust in human-robot collaborative tasks [Chen et al., 2018; Wang et al., 2016; Nikolaidis et al., 2017; Huang et al., 2018].

We first describe results from a human-subjects study (n = 32) in two domains—household object handling and autonomous driving—and show that inter-task trust transfer depends on perceived task similarity, task difficulty, and observed robot performance. Our findings are consistent over both domains, even though the robots and the contexts are different. Specifically, the household domain involves a Fetch robot navigating and handling various household objects. The driving domain involves a simulated autonomous vehicle performing driving and parking maneuvers.

Based on the findings, we formalize trust as a context-dependent latent dynamic function that changes with observations of robot performance across tasks. We focus on the representation and dynamics of this trust function and develop two specific differentiable models: (i) a Bayesian Gaussian process (GP) [Rasmussen and Williams, 2006] model and (ii) a connectionist model based on recent advances in deep learning. The GP model explicitly encodes a prior assumption that human trust evolves via Bayes rule. In contrast, the neural model is largely data-driven. Both models leverage latent task space representations learned using word vector descriptions of tasks, e.g., “Pick and place an apple”. Our Bayesian and neural models better predict self-reported human trust compared to existing approaches. From one perspective, the GP model extends the single global trust variable used in prior work [Chen et al., 2018; Xu and Dudek, 2015] to a collection of latent trust variables. To be clear, both models are computational analogues of trust, and neither model attempts to represent exact trust processes in the human brain. They offer conceptual frameworks for capturing the principles of trust formation and transfer.

2 Human Subjects Study
In this section, we summarize the key findings from our human subjects study. For more details, please refer to [Soh et al., 2018]. In brief, we find that human trust changes and transfers across tasks in a structured manner: observations of robot performance have a greater affect on the trust dynamics over similar tasks compared to dissimilar tasks.
Experimental design. Our experimental design is summarized 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. We used tasks in two domains:

- **Household** comprising two common categories of household tasks, i.e., picking and placing objects, and navigation in an indoor environment. We used a real-world Fetch robot to perform live in-lab task demonstrations.
- **Driving** involving a simulated autonomous vehicle (AV) performing lane merging and parking, with dynamic and static obstacles. Participants interacted with the simulation system via a Virtual Reality (VR) headset that presented a first-person viewpoint from the driver seat.

Note the same experiment protocol was conducted independently in each domain; having two separate domains enabled us to evaluate the robustness of our findings to different contexts.

Measures and procedure. The main dependent variable was subjective trust in the robot/agent $a$’s capability to perform specific tasks. Participants were first asked to indicate their subjective trust on three “tested tasks”. Participants were then randomly assigned to observe two tasks from a specific category and difficulty, and asked to indicate their trust in the robot to accomplish the observed tasks. Finally, participants were asked to re-indicate their trust on the three tested tasks. Participants indicated their degree of trust given task $x$ at time $t$, denoted as $r_{x,t}^a$, and we computed two derivative scores:

- **Trust Distance** $d_{r,t}(x,x') = |r_{x,t}^a - r_{x',t}^a|$, i.e., the 1-norm distance between scores for $x$ and $x'$ at time $t$.
- **Trust Change** $\Delta r_{x,t}^a(t_1,t_2) = |r_{x,t_1}^a - r_{x,t_2}^a|$, i.e., the 1-norm distance between the scores for $x$ at $t_1$ and $t_2$.

2.1 Key Results

We recruited 32 individuals (Mean age: 24.09 years, $SD = 2.37$, 46% female). For the driving domain, one partici-
3 Multi-Task Trust Models

The results from our human subjects study indicate that trust is relatively rich mental construct. Here, we present a computational model where trust is a latent dynamic function:
\[
\tau^a_t(x) : \mathbb{R}^d \rightarrow [0, 1]
\]
that maps task features, \( x \), to real-values indicating trustworthiness of the robot to perform the task. This functional view of trust captures trust differences across tasks, and can be extended to include other contexts, e.g., the current environment, robot properties, and observer characteristics.

To model trust dynamics, we propose a Markovian function \( g \) that updates trust,
\[
\tau^a_t = g(\tau^a_{t-1}, o^a_{t-1})
\]
where \( o^a_{t-1} = (x_{t-1}, e^a_{t-1}) \) is the observation of robot \( a \) performing a task with features \( x_{t-1} \) at time \( t-1 \) with performance outcome \( e^a_{t-1} \). The function \( g \) changes 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 \( e^a_{t-1} \in \{+1,-1\} \) indicating success and failure respectively, but differences in performance can be directly accommodated via “soft labels” \( e^a_{t-1} \in [-1,+1] \) without significant modifications to the following methods.

In this work, we propose and evaluate two different forms for \( \tau^a_t \) and \( g \): (i) a Bayesian approach where we model a probability distribution over latent functions via a Gaussian process, and (ii) a connectionist approach utilizing a recurrent neural network (RNN).

3.1 Bayesian Gaussian Process Trust Model

Our first model formalizes trust formation as a cognitive process in a rational Bayesian framework [Griffiths et al., 2009], i.e., the human is learning about the capabilities of the robot \( f^a \) by combining prior beliefs about the robot with evidence (observations of performance) via Bayes rule,
\[
p_t(f^a|o^a_{t-1}) = \frac{P(o^a_{t-1}|f^a, x_{t-1})p_{t-1}(f^a)}{\int P(o^a_{t-1}|f^a, x_{t-1})p_{t-1}(f^a)df^a},
\]
where \( p_t \) is the posterior distribution, \( P(o^a_{t-1}|f^a, x_{t-1}) \) is the likelihood of observing the robot performance \( o^a_{t-1} \) given the task \( x_{t-1} \) and latent function \( f^a \). The human estimates robot trustworthiness by integrating over the posterior:
\[
\tau^a_t(x_t) = \int P(c^a_{t-1}|f^a, x)p_t(f^a)df^a
\]
We place a Gaussian process (GP) prior over \( f^a \),
\[
p_0(f^a) = N(m(x), k(x, x'))
\]
where \( m(\cdot) \) is the prior mean function, and \( k(\cdot, \cdot) \) is the kernel or covariance function. In standard machine learning scenarios, GPs are often assumed to have a zero-mean prior, \( m(\cdot) = 0 \). However, as our human subject studies have shown, perceived task difficulty results in asymmetric trust transfer. As such, we use a linear prior mean function \( m(x) = \beta^T x \). In addition, we assume tasks to live on a low-dimensional manifold, i.e., a “psychological task space” and use a projection kernel:
\[
k(x, x') = \exp(-|(x - x')^\top M(x - x')|)
\]
trust data associated with one task and trained on the remaining tasks. In both experiments, we evaluated 5 different models: RNN: The neural trust model; PMGP: The GP trust model with prior mean; GP: A zero-mean Gaussian process trust model; LG: A linear Gaussian trust model similar to the linear Gaussian updates used in OPTImo [Xu and Dudek, 2015], and trained in the same manner as the RNN and GP models; and CT: A baseline model with constant trust.

In our evaluations, training involved learning the relevant model parameters, e.g., $\beta$ and $M$ for the GP. For each participant, each model was updated only with the tasks and outcomes that the participant observed. Prediction and evaluation was carried out on both pre- and post-update trust scores. All models were trained and tested using the data collected in our human subjects study. Preliminary cross-validation runs were used to ascertain good parameters (details in [Soh et al., 2018]). For task features, we used 50-dimensional GloVe word vectors [Pennington et al., 2014] computed from the task descriptions. Models were optimized via ADAM algorithm [Kingma and Ba, 2014] using the Bernoulli likelihood of observing the trust scores (as soft labels); when trust is unobserved, the models can be trained using observed human actions.

To mitigate significant differences across the folds (each participant/task split), we compared the methods using 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 $i$. $\text{MAE}_{\text{DfB}}$ is similarly defined. Our key results still hold when comparing NLL and MAE scores.

4.1 Results

Results for E1 are summarized in the boxplots shown in Fig. 3; we only show the $\text{NLL}_{\text{DfB}}$ measure as the $\text{MAE}_{\text{DfB}}$ was very similar. The RNN and PMGP models appear comparable, and outperform the baselines on both datasets. This suggests our models—which account for trust differences and task transfer—achieved better trust predictions on unseen participants (Q1).

Turning our attention to E2, we see that the RNN and PMGP once again outperform the other models (Fig. 4). Interestingly, the gap between them is larger on E2, with the RNN achieving lower scores than the PMGP. This suggests that the RNN learned a better mapping from the task feature space to the latent task space. Nevertheless, both models are able to make accurate trust predictions on unseen tasks (Q2).

Finally, to answer Q3, we examined the differences between the GP models. The PMGP model always achieved lower or similar scores to the GP model, suggesting that the difficulty modeling enabled better performance. While not conclusive evidence—the PMGP is more flexible due to the extra parameters—it does suggest the importance of accounting for the asymmetries resulting from prior beliefs.

5 Discussion

Human trust in automation is a large interdisciplinary research endeavor spanning multiple fields, including human-factor engineering, psychology, and human-robot interaction [Lee and Moray, 1992; Muir, 1994; Hancock et al., 2011; Chen et al., 2018; Soh et al., 2018]. Even so, crucial gaps remain in our understanding of when and how humans trust autonomous agents. In particular, how do we formalize trust, its representation and dynamics given variability among people, tasks, and robots? This work takes a key first step towards conceptualizing and formalizing computational models for predicting human trust across tasks. The following highlights the interesting new questions arising from our findings:

- **Combining Bayesian and neural trust models.** A natural question is whether we can formulate a “structured” trust update that combines the simplicity and interpretability of the Bayes update, while allowing for additional flexibility via a neural network. Preliminary experiments show such a hybrid model is able to achieve better prediction accuracy.

- **Trust in intent.** In this work, we limited our investigation to trust in the robot’s capabilities. However, it is also essential to examine trust in the robot’s “intention”, e.g., its policy [Huang et al., 2018] and decision-making process. In very recent work [Xie et al., 2019], we examined how human mental models of both these factors influence decisions to trust robots.

- **Trust-based decision making.** Our prior work [Chen et al., 2018] showed that decision-making (via a POMDP) using a human trust model results in policies that improved overall task performance. Integrating more complex models, such as the ones proposed in this work, into a decision-making framework for real-time trust inference and calibration remains a key open problem.

![Figure 3: NLL_{DfB} scores for experiment E1 with medians (lines) and means (triangles) shown. The proposed neural and Bayesian Trust models (RNN and PMGP) achieve better scores on previously unseen participants, than the alternative models that do not consider trust transfer.](image1)

![Figure 4: NLL_{DfB} scores for experiment E2 with medians (lines) and means (triangles) shown. Compared to existing models, the proposed methods—RNN and PMGP—are able to better predict trust on previously unseen tasks.](image2)
References


