# Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes

Shreyas Bhat
shreyasb@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Joseph B. Lyons
joseph.lyons.6@us.af.mil
Air Force Research Laboratory
Dayton, Ohio, USA

Cong Shi
congshi@bus.miami.edu
Miami Herbert Business School
Miami, Florida, USA

X. Jessie Yang
xijyang@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

## ABSTRACT

This paper examines the effect of real-time, personalized alignment of a robot's reward function to the human's values on trust and team performance. We present and compare three distinct robot interaction strategies: a non-learner strategy where the robot presumes the human's reward function mirrors its own; a non-adaptive-learner strategy in which the robot learns the human's reward function for trust estimation and human behavior modeling, but still optimizes its own reward function; and an adaptive-learner strategy in which the robot learns the human's reward function and adopts it as its own. Two human-subject experiments with a total number of $N = 54$ participants were conducted. In both experiments, the human-robot team searches for potential threats in a town. The team sequentially goes through search sites to look for threats. We model the interaction between the human and the robot as a trust-aware Markov Decision Process (trust-aware MDP) and use Bayesian Inverse Reinforcement Learning (IRL) to estimate the reward weights of the human as they interact with the robot. In Experiment 1, we start our learning algorithm with an informed prior of the human's values/goals. In Experiment 2, we start the learning algorithm with an uninformed prior. Results indicate that when starting with a good informed prior, personalized value alignment does not seem to benefit trust or team performance. On the other hand, when an informed prior is unavailable, alignment to the human's values leads to high trust and higher perceived performance while maintaining the same objective team performance.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computer systems organization** → **Robotic autonomy**.

## KEYWORDS

Human-robot teaming, trust-aware decision-making, value-alignment



*HRI '24, March 11–14, 2024, Boulder, CO, USA*



ACM ISBN 979-8-4007-0322-5/24/03…$15.00
https://doi.org/10.1145/3610977.3634921

**ACM Reference Format:**
Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2024. Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634921

## 1 INTRODUCTION

Robots are increasingly becoming an integral part of our daily lives, marking their presence across varied domains, including healthcare, manufacturing, education, and home assistance, to name a few. As this integration deepens, robots are no longer perceived as tools performing isolated tasks; they are evolving as collaborative partners working with humans. In this human-robot partnership, research into trust between humans and robots becomes increasingly important [6, 10, 35, 42]. Without proper trust, the potential of human-robot teams remains unrealized.

A considerable amount of research has been devoted to developing robots exhibiting trustworthy behaviors, as well as investigating methods for predicting and managing the human's trust in the robot. For instance, one specific area of research focuses on providing explanations of the robots' behaviors [14, 28, 29, 37, 38] typically leading to higher perceived trustworthiness and, subsequently, trust in the robot. Other research directions include the developing real-time trust prediction algorithms [18, 19, 36, 39, 41], modeling trust dynamics [12, 40, 41], developing trust repair strategies [15, 22], and developing trust-aware planning [2, 5, 8, 9, 33, 44].

More recently, the idea of value/goal alignment – aligning the values/goals of robots with those of humans, has garnered significant attention, with the assumption that such alignment would benefit human-robot interaction [34, 43]. Recent literature in value/goal alignment is primarily focused on enabling the autonomous or robotic agent to learn the human's values/goals through preferences [7, 11, 20] or demonstrations [3, 16, 20]. However, there is a lack of research empirically examining and quantifying the effects of alignment on human-robot interaction processes and outcomes. Yet, there are at least three reasons to suggest that such alignment could be beneficial. First, prior research has illustrated that agent adaptation to humans can enhance performance [4, 26]. Second, agent adaptation could be viewed as the agent being responsive

to the human and may, in turn, increase human trust in the agent and enhance team performance [25]. Third, value alignment not only could facilitate trust establishment and enhance team performance, but it is also important for ensuring that machine partners are morally acceptable [24].

This study investigates the effect of real-time, personalized alignment of a robot's reward function to the human's values on trust and human-robot team performance through two human-subject studies. We model the interaction between the human and the robot as a trust-aware Markov Decision Process (trust-aware MDP) and use Bayesian Inverse Reinforcement Learning to estimate the reward weights of the human as they interact with the robot. We compare three types of robot interaction strategies: (1) the non-learner strategy, where the robot presumes the human's reward function mirrors its own; (2) a non-adaptive-learner strategy, in which the robot learns the human's reward function for trust estimation and human behavior modeling, but still optimizes its own reward function; and (3) an adaptive-learner strategy where the robot learns the human's reward function and aligns to it. In addition, we employ different initial conditions for the IRL learning algorithm in the two experiments, one with an informed prior and the other with an uninformed prior.

Results indicate that when starting with an informed prior, personalized alignment to values does not seem to benefit trust or team performance. On the other hand, when an informed prior is unavailable, aligning to the human's values leads to higher trust, agreement, and reliance intentions while maintaining the same objective team performance. To the best of our knowledge, this is one of the few studies, if not the only, that provides empirical evidence for the benefits of value alignment.

The rest of the paper is organized as follows: Section 2 gives an overview of related work that our study builds upon. Section 3 details the human-robot team task and formulates our problem as a trust-aware Markov Decision Process (trust-aware MDP). Section 4 details the human-subjects experiment. Section 5 discusses major results and their implications. Finally, section 6 concludes our study and discusses limitations and future work.

## 2 RELATED WORK

Our study is motivated by two bodies of research. The first is using Inverse Reinforcement Learning (IRL) [32] to learn from human demonstrations and/or preferences to guide the robot's behavior. The second deals with trust-aware planning and quantitative modeling of trust, the goal of which is to estimate the human's trust level during interaction and to use the estimated value of trust to plan behaviors for the robot.

### 2.1 Value Alignment

Over the past few years, the problem of aligning the values/goals of the robot to those of its human teammate has been studied in detail in human-robot teaming literature [16, 20, 30, 43].

A bidirectional value alignment problem is studied by Yuan et al. [43]. In their study, the human knows the true reward function and behaves accordingly while interacting as a supervisor to a group of worker robots. The robots try to learn this true reward function through correctional inputs to their behavior from the human. The human, on the other hand, tries to update her belief on the robot's belief of the true reward function and inputs corrections to their behavior accordingly. They compare the degree of alignment of human estimates of the robot's value function and the degree of alignment of the robot's value function to the true value function. Their results reveal evidence of a bidirectional value learning behavior from the human and the robot.

Hadfield-Menell et al. [20] formally define the value alignment problem as cooperative inverse reinforcement learning (CIRL). They show that the more traditional framework of apprenticeship learning can be formulated as a CIRL game. Their results indicate that the human acting optimally in isolation may not be an effective way to teach the robot. They show that under the CIRL formulation behaviors such as active teaching, active learning, and communicative actions become optimal.

Fisac et al. [16] discusses a solution to the CIRL game based on established models of cognition and theory of mind. Under this solution, the human thinks pedagogically, choosing actions that give the most information to the robot about the underlying reward function. The robot, in turn, expects this behavior and acts pragmatically on the human's behavior. This enables the robot to learn the reward function quickly and efficiently.

Christiano et al. [11] proposes an algorithm to learn from human preferences and shows that it can be used to "solve" reinforcement learning tasks in which the robot's goal is to minimize cost to reach a goal state. They show that this can be done for complex tasks within an hour of the human's time. Additionally, they show that by incorporating human preferences, the robot can learn more efficiently than using traditional deep reinforcement learning methods.

Milli et al. [30] compare a robot that completely abides by the human's literal order with a robot that instead behaves according to its estimate of the human's underlying preferences. They use simulations to compare how much more reward the human would get if the robot directly followed the human's orders vs if the robot used an estimate of the human's preferences. Their results indicate that 1) when a human is not rational, a robot should not directly obey their commands, 2) The optimal robot obeys only optimal commands from the human, and uses the estimate of the posterior mean on the reward features to drive its behavior otherwise.

There are two main differences between our work and prior literature in value alignment. Firstly, most prior work in value alignment deals with a human-supervisor robot-worker scenario (the robot performs some task and the human is free to interrupt the task if they see any unexpected behavior from the robot) or scenarios where the human demonstrates ideal behaviors to the robot. In our case, however, we want to predict and use the probability of the human accepting or rejecting recommendations from the robot (i.e., robot-recommender human-follower scenario). This calls for the embedded trust dynamics and human behavior model, which is absent in most previous works in this area. Secondly, in our case, there is no *true* reward function: the human and the robot have their own reward functions, and we want to see the effect of aligning/not aligning the robot's reward function with that of the human on trust and team performance. Such studies have not been done previously, according to the best of our knowledge.

## 2.2 Trust-Aware Decision Making

In recent years, there has been substantial research in developing trust-aware decision-making algorithms for robots that work with humans in collaborative tasks. These works model the human's trust in the robot explicitly in the decision-making framework and leverage the use of quantitative trust and trust-behavior models.

Guo et al. [17] and Bhat et al. [5] present the use of the Beta distribution trust model [19] in a sequential decision-making scenario for a human-robot team. They use this model to define a trust-aware Markov Decision Process, which gives optimal actions for the robot considering the human's trust and subsequent behavior.

Akash et al. [1] model the human-robot interaction as a Partially Observable Markov Decision Process (POMDP) with human's trust and workload as the states. A solution to their formulation is presented in [2] which gives optimal level of transparency for the robot's interface depending on the human's level of trust.

Chen et al. [9] propose a trust-POMDP that is solved to generate optimal trust-based policies for the robot. They demonstrate it in a human-subject experiment involving pick-and-place tasks for a human-robotic arm team. The robot chooses an "easy" object to pick and place to build trust and moves to harder objects when trust is high to minimize interruptions from the human.

Zahedi et al. [44] present a trust dynamics model and a meta-MDP framework that chooses a robot's behavior depending on the level of trust of the human. They analyze the case where the human's model of the environment may be false leading to sub-optimal trustworthy policies and untrustworhy optimal policies. Their framework generates an optimal policy for the robot that chooses the trustworthy sub-optimal policies when trust is low in order to increase it and chooses the untrustworthy optimal policy when trust is high enough to improve the team's performance.

A majority of these works model the human-robot interaction as a reward-maximization problem with a reward function that is known to the team. Our work differs in this respect; we offer participants the autonomy to formulate their own reward functions, providing them only with a broad understanding of the team's objectives. Subsequently, we explore how discrepancies in reward functions between humans and robots influence trust and overall team performance.

# 3 PROBLEM FORMULATION

This section describes the task for the human-robot team and the mathematical formulation of the interaction.

## 3.1 Human-Robot Teaming Task

We designed a scenario in which the human-robot team performs a search for potential threats in a town. The team sequentially goes through search sites to look for threats. At each site, the team is given a probability of threat presence inside the site via a scan of the site by a drone. The robot additionally, has some prior information about the probability of threat presence at all of the search sites. This prior information is unknown to the human, thus creating interdependence between the human and drone. After getting the updated probability of threat presence, the robot solves the trust-aware MDP to generate a recommendation for the human. It can either recommend the human to use or not use an armored robot for protection from threats. Encountering a threat without protection from the armored robot will result in injury to the human. On the other hand, using the armored robot takes extra time since it takes some time to deploy and move the armored robot to the search site. The goal of the team is to finish the search mission as quickly as possible while also maintaining the soldier's health level. Thus, a two-fold objective arises with conflicting sub-goals: To save time you must take risks, and if you want to avoid risks, you must sacrifice precious mission time.

## 3.2 Markov Decision Process

We model the interaction between the human and the robot as a trust-aware Markov Decision Process (trust-aware MDP), which is a tuple of the form $(S, A, T, R, H)$, where $S$ is a set of states one of which is the trust of the human on the robot, $A$ is a finite set of actions, $T$ is the transition function, $R$ is a reward function and $H$ is an embedded human trust-behavior model, which gives the probabilities of the human choosing a certain action given the action chosen by the robot, their level of trust, etc. Below we provide details of our MDP formulation.

*3.2.1 States.* We use the level of trust $t \in [0, 1]$ as the state in our trust-aware MDP formulation. The dynamics of trust are thus described by our transition function.

*3.2.2 Actions.* At any site, the recommender robot has two choices of action. It can either recommend to use or not use the armored robot. These are represented by the binary actions $a^r = 1$ and $a^r = 0$, respectively. Thus, our action set is $A = \{0, 1\}$. After receiving a recommendation, the human chooses action $a^h$ from the same action set.

*3.2.3 Reward Function.* The rewards for both agents (the human and the robot) are a weighted sum of the negative cost of losing health and losing time. The weights for these costs can be different for the robot and the human. In general, for agent $o \in \{h, r\}$, the reward function can be written as,

$$R^o(D, a) = -w_h^o h(D, a) - w_c^o c(a). \tag{1}$$

Here, $D$ is a random variable representing the presence of threat inside a search site, $a$ is the action chosen by the human to implement, $o \in \{h, r\}$ represents the agent, either the human $h$ or the robot $r$. Note that $h(D, a)$ is a function giving the health loss cost and $c(a)$ is a function giving the time loss cost.

*3.2.4 Transition Function.* The transition function gives the dynamics of trust as the human interacts with the robot. We use the model from Bhat et al. [5] which models trust as a random variable following the Beta distribution based on personalized parameters $(\alpha_0, \beta_0, v^s, v^f)$. Here, $\alpha_0, \beta_0$ are a measure of the initial trust of the human, while $v^s$ and $v^f$ control the effect of successes and failures on trust respectively. More specifically, the trust level after $i$ interactions is given by

$$t_i \sim Beta(\alpha_i, \beta_i), \tag{2}$$

where the parameters are updated through

$$\alpha_i = \alpha_0 + \sum_{j=1}^{i} p_j v^s, \tag{3}$$

$$\beta_i = \beta_0 + \sum_{j=1}^{i} (1 - p_j) v^f. \tag{4}$$

Here, $i$ is the number of interactions completed between the human and the robot, $t_i$ is the current level of trust, and $p_j$ is the realization of the random variable performance ($P_j$) of the recommender robot at the $j^{th}$ interaction. It is defined below.

$$P_j = \begin{cases} 1, & \text{if } R_j^h(a_j^r) \geq R_j^h(1 - a_j^r), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here, $R_j^h(a_j^r)$ is the reward for the human for choosing the recommended action ($a_j^r$) at the $j^{th}$ interaction and $R^h(1 - a_j^r)$ is the reward for choosing the other action ($1 - a_j^r$). Thus, essentially, the transition function shifts the distribution to the right if the recommend action led to a better immediate reward to the human. Otherwise, it shifts the distribution to the left.

*3.2.5 Human Trust-Behavior Model.* A human trust-behavior model gives the probabilities of a human choosing an action, given the robot's action, their trust level, and other factors such as the human's goals/values. In our study, we make use of the Bounded Rationality Disuse Model of human trust-behavior. This model states that the human chooses the recommended action with a probability equal to the human's current level of trust. If the human chooses to ignore the recommendation, s/he will choose an action according to the bounded rationality model of human behavior. That is, the human will choose an action with a probability that is proportional to the exponential of the expected reward the human receives with that action. Mathematically,

$$P(a_i^h = a | a_i^r = a) = t_i + (1 - t_i) q_a, \tag{6}$$

$$P(a_i^h = 1 - a | a_i^r = a) = (1 - t_i)(1 - q_a). \tag{7}$$

where $t_i$ is the human's level of trust at the $i^{th}$ interaction and $q_a$ is the probability of choosing action $a \in \{0, 1\}$ under the bounded rationality model [16, 30, 45]. It is given by,

$$q_a = \frac{\exp(\kappa E[R_i^h(a)])}{\sum_{a' \in \{0,1\}} \exp(\kappa E[R_i^h(a')])}. \tag{8}$$

Here, $\kappa$ is called the rationality coefficient of the human, with a higher value indicating a more rational human, and a value of 0 indicating a human that chooses an action at random. Note that this model can be easily extended to the case where multiple actions are possible for the human-robot team. We will just need to sum over all actions in the denominator to get the probabilities.

## 3.3 Bayesian Inverse Reinforcement Learning

We use Bayesian IRL to estimate the reward weights of the human as they interact with the recommender robot. This is done by maintaining a distribution on the possible reward weights and updating it using Bayes' rule after observing the human's selected action. More precisely, if $b_i(w)$ is the belief distribution on the reward weights before the $i^{th}$ interaction, the distribution after the $i^{th}$ interaction, $b_{i+1}(w)$ is given by,

$$b_{i+1}(w) \propto \begin{cases} P(a_i^h = a_i^r | a_i^r) b_i(w), & \text{if } a_i^h = a_i^r, \\ P(a_i^h = 1 - a_i^r | a_i^r) b_i(w), & \text{otherwise.} \end{cases} \tag{9}$$

In our formulation, we only learn a distribution over the health reward weight of the human, $w_h^h$, and assume that the time reward weight is defined by $w_c^h := 1 - w_h^h$. We use the mean of the learnt distribution as an estimate of the human's health reward weight. The Bayesian IRL algorithm requires a prior distribution $b_0(w)$ to get started. We ran the algorithm starting with a uniform prior on previously collected data [5] and generated an "average" distribution that can represent the weights for the general population. In Experiment 1, we use this informed prior for $b_0(w)$ and in Experiment 2, we use the uniform distribution for $b_0(w)$ to simulate a case where we lack data to set an informed prior.

# 4 EXPERIMENT

This section provides details about the testbed and the human-subject experiments. The experiments complied with the American Psychological Association code of ethics and were approved by the Institutional Review Board at the University of Michigan.

## 4.1 Testbed

We developed a 3D testbed using the Unreal Engine game development platform. A soldier moves with an autonomous drone in a town to search for threats (armed gunmen). Before entering a site, the drone scans the site and reports the chance of threat presence inside the site (Fig. 1). Then, the participant is presented with the average time taken to search a site with and without the armored robot to aid their decision. Finally, the participants are given the recommendation generated by the intelligent agent. If the participant chooses to use the armored robot, the armored robot is shown to be moving towards the site. This deployment of the armored robot takes about 15 seconds. If the participant chooses to not use the armored robot, they enter the site directly without any time loss. In case a threat is encountered without protection from the armored robot, the participant loses 5 points of health. The participant does not lose any health if there is no threat or if there is a threat but the participant chose to use the armored robot. After exiting each house, the participants are asked to adjust a slider to give feedback on their level of trust on the agent's recommendations (Fig. 2). The feedback slider shows the threat level, the recommendation, the participant's choice, the presence of threat, and the time it took to search the site to help the participants in assessing their trust.

## 4.2 Participants

We collected data from a total of 54 participants, 30 of which participated in experiment 1 (Age: Mean 22.6 years, *SD* 3.6, 14 Female) and 24 participated in experiment 2 (Age: Mean 21.4 years, *SD* 2.3, 12 Female). All participants were students from the University of Michigan.

## 4.3 Experiment Design

We designed three interaction strategies for the intelligent agent:

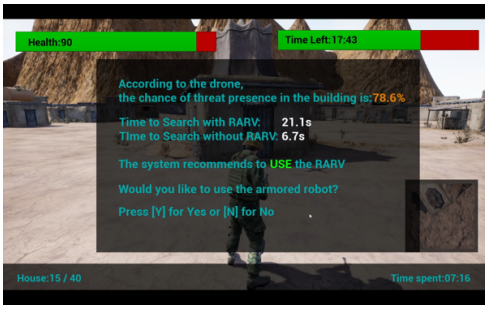


**Figure 1: The recommendation interface**

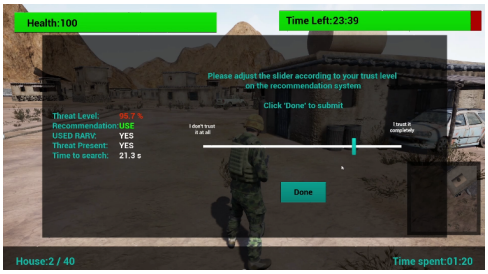


**Figure 2: The trust feedback slider used to get feedback from the participants after every search site. The mission timer is paused when the slider is shown to let the participants take their time in adjusting their trust.**

- **Non-learner:** The intelligent agent does not learn the reward weights of the human. It assumes that the human and the intelligent agent share the same reward weights and uses these for recommendation success assessment, trust updating, human behavior modeling, and MDP optimization.
- **Non-adaptive learner:** The intelligent agent learns personalized reward weights for each human it interacts with. It only uses these learned weights for recommendation success assessment, trust updating, and human behavior modeling. It still optimizes the MDP with its own fixed reward weights.
- **Adaptive learner:** The intelligent agent learns personalized reward weights for each human. It uses them for recommendation success assessment, trust updating, and human behavior modeling, and also optimizes the MDP based on these reward weights.

We employed a within-subjects design. Each participant completed three missions. In each mission, they interacted with an intelligent agent using one of the interaction strategies. To minimize potential order effects, a $3 \times 3$ Latin Square design was used.

## 4.4 Measures

*4.4.1 Pre-experiment Measures.* Prior to the experiment, participants filled in a demographic survey indicating their age, gender, academic department, nationality, frequency and skill of playing video games, and familiarity with AI/ML algorithms. Participants also filled in questionnaires about their personality, propensity to trust autonomy, and decision-making style.

*4.4.2 Pre-mission Measures.* Before each of the three missions, participants rated their preferences.

- **Task Preference:** Before the beginning of each mission, we ask the participants to rate their preference between saving health and saving time by moving a slider between these two objectives, showing their relative importance.

*4.4.3 In Experiment Measures.* After each site's search was completed (i.e., every trial), the participants were asked to report their level of trust in the intelligent agent, $t_i$ (see fig. 2 for the exact question asked during the interaction). The slider values were between 0 and 100 with a step of 2 points. Additionally, for every trial, we measured whether the participants agreed with the intelligent agent. With the trial-based data, we measured the following:

- **Average Trust:** This was calculated as the empirical mean trust $\frac{1}{M}\sum_{i=1}^{M} t_i$.
- **End-of-mission Trust:** This was the participant's self-reported trust after the last trial, $t_M$.
- **Number of Agreement:** This was computed as the number of times the participant chose the recommended action.

Note that $M = 40$ is the number of sites in a mission.

*4.4.4 Post-mission Measures.* After every mission, participants filled out a post-mission survey gauging the following items.

- **Post-mission trust questionnaire:** This was measured using Muir's trust questionnaire [31]. It has 9 questions, each with a slider range between 0 and 100.
- **Post-mission Reliance Intentions:** This was measured using the scale developed in Lyons and Guznov [27]. We used 6 of the 10 items that were relevant for this task. Each item was rated on a 7-point Likert scale.
- **Workload:** Workload was measured using the NASA TLX scale [21]. We used 5 of the 6 dimensions as our experiment involved minimal physical demand. Each item was measured using a slider ranging from very low to very high.
- **Performance:** We computed the team performance by a weighted sum of the percentage health remaining of the soldier and the percentage time remaining in the mission.

$$\text{Performance} = \hat{w}_h^h \cdot (\%h_M) + \hat{w}_c^h \cdot (100 - \%c_M). \tag{10}$$

where $\hat{w}_h^h$ and $\hat{w}_c^h := 1 - \hat{w}_h^h$ are the reported preferences by the participant before beginning the mission, $\%h_M$ is the percent health remaining and $\%c_M$ is the percent time spent at the end of the mission. This metric allows us to convert the two conflicting objectives with different units of measurement into one unified scale. The higher the value, the better the team performance.

## 5 RESULTS

This section summarizes our results and discusses the implications. Table 1 tabulates the results from the two experiments. Repeated measures analyses of variance (ANOVAs) were conducted to compare the three interaction strategies. Greenhouse-Geisser corrections to the degrees of freedom were made whenever a measure failed Mauchly's test of sphericity. In Experiment 1, we initiated

the IRL learning algorithm with an informed prior. In Experiment 2, the learning algorithm started with an uninformed prior.

## 5.1 Trust, Agreement, and Reliance Intention

*5.1.1 Experiment 1: with informed prior.* We observed no significant difference between the three strategies in average trust ($F(2, 58) = 0.308$, $p = 0.736$), end-of-mission trust ($F(2, 58) = 1.192$, $p = 0.311$), and the Muir's trust scale ($F(2, 58) = 1.550, p = 0.221$).

Additionally, there was no significant difference in the number of agreements ($F(2, 58) = 0.755$, $p = 0.475$) across the three strategies. However, there was a significant difference in reliance intentions ($F(1.543, 44.737)$, $p = 0.031$) (Fig. 3). Pairwise comparisons with Bonferroni adjustments revealed a lower intent to rely on the adaptive learner strategy than the non-learner strategy ($p = 0.012$).

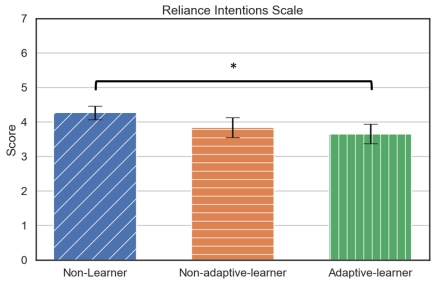


**Figure 3: Exp 1 – Post-mission reliance intentions**

*5.1.2 Experiment 2: with uninformed prior.* Figs. 4, 5, 6 show the comparisons of the three strategies in trust. Repeated measures ANOVA revealed significant differences between the three strategies in average trust ($F(2, 46) = 14.161$, $p < 0.001$), end-of-mission trust ($F(2, 46)) = 12.736$, $p < 0.001$, and Muir's trust scale ($F(1.586, 36.473) = 16.3$, $p < 0.001$). Pairwise comparisons with Bonferroni adjustments revealed that the adaptive-learner strategy led to higher average trust, end-of-mission trust, and post-mission trust compared to the non-learner strategy ($p < 0.001$, $p = 0.001$ and $p < 0.001$, respectively) and compared to the non-adaptive learner strategy ($p = 0.003$, $p < 0.001$, $p < 0.001$, respectively).

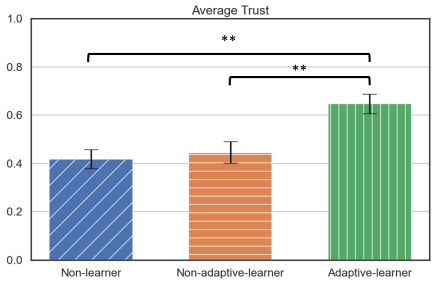


**Figure 4: Exp 2– Average trust $\frac{1}{M}\sum_{i=1}^{M} t_i$**

Regarding the number of agreements (Fig. 7), there was a significant difference among the three strategies ($F(1.584, 36.435) = 25.829$, $p < 0.001$). Post-hoc analysis showed that there was a significant difference between the non-learner and the adaptive-learner strategies ($p < 0.001$) and between the non-adaptive-learner and adaptive-learner strategies ($p < 0.001$).

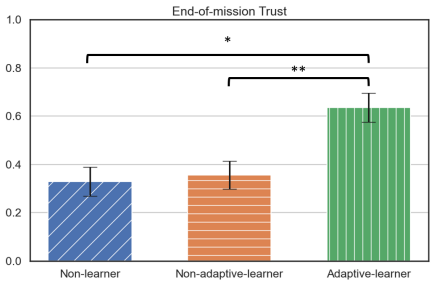


**Figure 5: Exp 2– End-of-mission trust $t_M$**

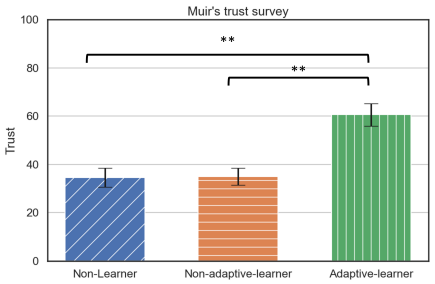


**Figure 6: Exp 2– Post-mission trust questionnaire**

Comparing reliance intentions (Fig. 8), there was a significant difference between the three strategies ($F(2, 46) = 13.691, p < 0.001$), with the adaptive-learner strategy rated higher than the non-learner strategy ($p < 0.001$) and the non-adaptive-learner strategy ($p = 0.004$).

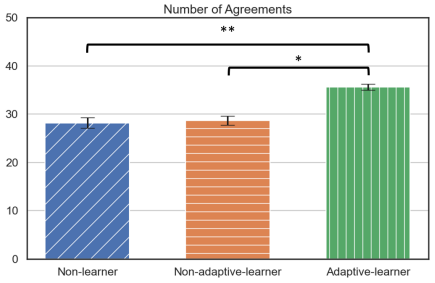


**Figure 7: Exp 2– Number of agreements**

## 5.2 Performance

*5.2.1 Experiment 1: with informed prior.* Even though there seemed to be a decreasing performance trend from non-learner to non-adaptive learner and to adaptive learner (i.e., 61.47 ± 18.12, 60.25 ±

**Table 1: Mean and standard deviation (SD) of dependent measures for the three interaction strategies in Experiments 1 and 2**

| **Experiment 1: with informed prior** | | | |
|---|---|---|---|
| | **Non-learner (Mean±SD)** | **Non-adaptive learner (Mean±SD)** | **Adaptive-learner (Mean±SD)** |
| Average trust $\frac{1}{M}\sum_{i=1}^{M} t_i$ | 0.73 ± 0.16 | 0.71 ± 0.20 | 0.72 ± 0.14 |
| End-of-mission trust $t_M$ | 0.81 ± 0.18 | 0.76 ± 0.21 | 0.76 ± 0.16 |
| Muir's trust questionnaire | 74.20 ± 14.80 | 71.09 ± 23.17 | 68.28 ± 17.4 |
| Number of agreements | 36.03 ± 3.72 | 35.27 ± 4.68 | 36.00 ± 2.60 |
| Reliance intentions scale * | 4.34 ± 1.01 | 4.10 ± 1.35 | 3.70 ± 1.27 |
| Workload | 36.87 ± 18.04 | 34.45 ± 17.02 | 39.03 ± 16.85 |
| Performance | 61.47 ± 18.12 | 60.25 ± 17.03 | 55.83 ± 20.39 |
| **Experiment 2: with uninformed prior** | | | |
| | **Non-learner (Mean±SD)** | **Non-adaptive learner (Mean±SD)** | **Adaptive-learner (Mean±SD)** |
| Average trust $\frac{1}{M}\sum_{i=1}^{M} t_i$ * | 0.42 ± 0.20 | 0.45 ± 0.22 | 0.65 ± 0.20 |
| End-of-mission trust $t_M$ * | 0.33 ± 0.30 | 0.35 ± 0.29 | 0.64 ± 0.30 |
| Muir's trust questionnaire * | 34.51 ± 19.79 | 34.97 ± 17.41 | 60.57 ± 23.71 |
| Number of agreements * | 28.17 ± 5.49 | 28.67 ± 4.65 | 35.62 ± 2.93 |
| Reliance intentions scale * | 2.16 ± 1.22 | 2.38 ± 1.16 | 3.58 ± 1.32 |
| Workload * | 38.18 ± 13.79 | 39.82 ± 14.94 | 31.13 ± 10.64 |
| Performance | 45.86 ± 16.20 | 49.11 ± 14.25 | 51.60 ± 18.68 |

$* - p < 0.05$

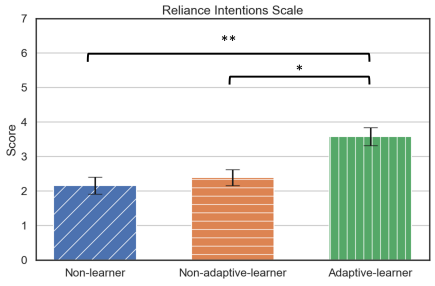


**Figure 8: Exp 2– Post-mission reliance intentions**

17.03, and 55.83 ± 20.39) the trend did not reach statistical significance ($F(2, 58) = 2.067,\ p = 0.136$).

*5.2.2 Experiment 2: with uninformed prior.* There seemed to be an upward trend from non-learner to non-adaptive learner and to adaptive-learner. Unfortunately, it did not reach significance.

## 5.3 Workload

*5.3.1 Experiment 1: with informed prior.* Comparing the average workload across the three interaction strategies showed no significant difference ($F(2, 58) = 2.634,\ p = 0.089$) (Table 1). Additionally, repeated measures ANOVAs did not reveal any significant differences between the three strategies in any of the dimensions.

*5.3.2 Experiment 2: with uninformed prior.* Comparing the average workload (1) across the three interaction strategies showed a significant difference ($F(2, 46) = 10872,\ p < 0.001$). Figure 9 shows the participants' responses on each dimension. There were significant differences between the three strategies in performance ($F(2, 46) = 5.443,\ p = 0.008$), effort ($F(2, 46) = 4.252,\ p = 0.02$), and frustration ($F(2, 46) = 5.454,\ p = 0.007$). Pairwise comparisons with Bonferroni adjustments showed that the adaptive-learner strategy led to higher perceived performance compared to the non-adaptive learner ($p = 0.037$) and to the non-learner ($p = 0.044$) strategies and led to lower frustration compared to the non-adaptive learner ($p = 0.032$) and to the non-learner ($p = 0.017$) strategies.

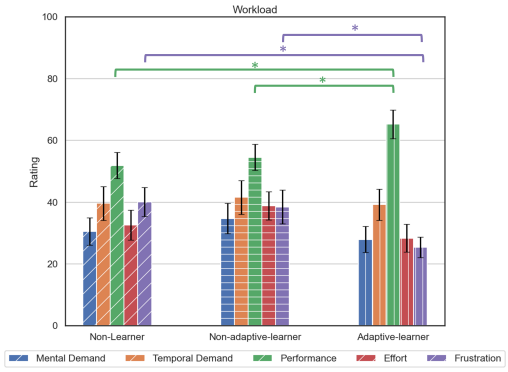


**Figure 9: Exp 2: Responses on the NASA TLX scale**

## 6 DISCUSSION AND CONCLUSION

In this study, we developed and compared three robot interaction strategies: the non-learner strategy, the non-adaptive-learner strategy, and the adaptive-learner strategy, with and without an informed prior for the IRL learning algorithm. We focused on evaluating their influence on various human trust in the robot, agreement, reliance, workload, and team performance.

### 6.1 Value Alignment with an Informed Prior

One critical insight from the research is the observed "uniformity" across all three interaction strategies when the IRL algorithm is initiated with an informed prior. In Experiment 1, the informed prior was calculated by training on a previously collected dataset using the same testbed and with participants from the same population as this study. Therefore, the calculated prior was considered an accurate estimation of the participants' true values/goals. The resulting prior was realistically skewed toward saving health rather than saving time. With this accurate informed prior, the three interaction strategies were more-or-less indistinguishable. Among all the dependent measures, including trust, agreement, reliance intentions, performance, and workload, we observed only one significant difference in reliance intentions. This large uniformity across the three strategies could be because, with an accurate informed prior on human values, there is no room for significant alignment to be done by the adaptive-learner strategy. This highlights that when a robot's reward function is already closely aligned with that of the human *a-priori*, adaptive strategies may exhibit negligible benefits.

Out of expectation, we observe a significant difference in post-mission reliance intentions, that participants were willing to rely on the non-learner more than the adaptive-learner strategies. This result could have been because any alignment from the adaptive-learner strategy, although very limited, could be regarded as a lack of predictability. As pointed out in research on teamwork, the ability to predict the actions of one's teammate is vital [13, 23].

### 6.2 Value Alignment with an Uninformed Prior

In Experiment 2, when the IRL learning algorithm was initiated with an uninformed uniform prior, we observed significant benefits of value alignment. The adaptive-learner strategy led to significantly higher trust, high agreement, and reliance intentions. In addition, participants had higher perceived performance and lower frustration interacting with the adaptive learner, while the team performance was maintained. Thus, in the absence of a good initial prior, our adaptive framework can be used to build trust in the robot. This scenario underlines the importance of adaptive strategies in real life where a good *a-priori* estimation of the human's values/goals is oftentimes unavailable, offering an approach to building trust while maintaining performance.

The results from Milli et al. [30] showed that robots should not be completely obedient to humans who are not acting rationally. Instead, when interacting with such humans, the robot should use its estimation of the human's underlying reward function. We extend this by providing a human-subjects study showing that personalized value alignment is only beneficial when a good prior on the human's reward weights is unavailable

### 6.3 Implications for a broader HRI context

The algorithmic focus of existing work on value alignment [16, 20, 43] has one implicit assumption: aligning a robot's values with a human's is beneficial. Our study is the first attempt to examine whether and to what extent such alignment can benefit trust, workload, and team performance. We show that personalized value alignment is beneficial only when an informed prior is unavailable. The implications of our study should extend to other real-life HRI scenarios involving conflicting objectives. For instance, a rehabilitation robot must balance a patient's pain tolerance with long-term health goals when assigning the appropriate level of exercise.

Our study involved a relatively homogeneous participant group, leading to the calculation of a fairly accurate informed prior. However, achieving such accuracy in real-world settings with demographically diverse individuals is more challenging. In such cases, aligning a robot's values with those of individual human users becomes essential. The benefits obtained from value alignment are key for the acceptance and adoption of robots in homes and workplaces, highlighting the need for adaptable strategies in HRI design.

Further, the idea of incorporating a layer of trust in the decision-making system of an intelligent agent trying to align its values to that of the human user is an interesting area to explore in other HRI domains like shared control, social robotics, etc.

### 6.4 Limitations and Future Research

The results of our study should be seen in light of the following limitations. First, we provide a demonstration in the case when there are only two components in the team's reward function. Therefore, we only need to learn the human's preference for one of the two components and can ascertain their relative preference between the two objectives. Our formulation, however, can readily be extended to the case where there are more than two objectives in the team's reward function, with additional computations required to learn and maintain a distribution over each reward weight.

Second, our simulated scenario consists of binary actions. Judging the performance of the recommendations is fairly easy in this case, since we only need to compare the rewards earned for these two actions. In case more than two actions are available, this assessment becomes more difficult. Thus, although the human trust-behavior model can be readily extended to such a case of multiple actions, extending the trust dynamics model is challenging and is an interesting avenue for future research.

Third, in our scenario, there is an expected skewness among the general population to be more concerned about saving health. It would be interesting to study cases where the two objectives are more balanced, resulting in a more balanced informed prior. In such cases, personalized adaptation may still be beneficial.

Finally, our scenario, which entails a trade-off between "saving health" and "saving time", and the decision to use or not use an armored robot, is informed by the complex decision-making scenarios in real-life HRI contexts, such as DARPA's SQUAD-X program in which individuals receive recommendations from air and ground robots for various tasks. While our scenario offers insights into these types of decisions, we recognize it as a simplified representation of situations where decisions involve numerous objectives, a variety of recommendations, and possible actions. Therefore, further research is essential to determine the applicability of our findings in more complex, real-world environments and to validate the robustness of our conclusions in diverse and dynamic HRI settings.

## ACKNOWLEDGMENTS

This work was supported by the Air Force Office of Scientific Research under Grant FA9550-20-1-0406.

# REFERENCES


[1] Kumar Akash, Katelyn Polson, Tahira Reid, and Neera Jain. 2019. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part I: Human Trust and Workload Model. *IFAC-PapersOnLine* 51, 34 (2019), 315–321. 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018.

[2] Kumar Akash, Tahira Reid, and Neera Jain. 2019. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part II: Control Design and Synthesis. *IFAC-PapersOnLine* 51, 34 (2019), 322–328. 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018.

[3] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (Aug. 2021), 103500. https://doi.org/10.1016/j.artint.2021.103500

[4] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2020. Context-Adaptive Management of Drivers' Trust in Automated Vehicles. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6908–6915. https://doi.org/10.1109/LRA.2020.3025736

[5] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2022. Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8815–8822. https://doi.org/10.1109/LRA.2022.3188902

[6] Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, and Peter A. Hancock. 2012. Human-robot interaction: Developing trust in robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 109–110. https://doi.org/10.1145/2157689.2157709

[7] Erdem Bıyık and Dorsa Sadigh. 2018. Batch Active Preference-Based Learning of Reward Functions. arXiv:1810.04303 [cs.LG]

[8] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with Trust for Human-Robot Collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 307–315. https://doi.org/10.1145/3171221.3171264

[9] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2020. Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *J. Hum.-Robot Interact.* 9, 2, Article 9 (jan 2020), 23 pages. https://doi.org/10.1145/3359616

[10] Erin K. Chiou and John D. Lee. 2023. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors* 65, 1 (2023), 137–165. https://doi.org/10.1177/00187208211009995 arXiv:https://doi.org/10.1177/00187208211009995 PMID: 33906505.

[11] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML]

[12] Myke C. Cohen, Mustafa Demir, Erin K. Chiou, and Nancy J. Cooke. 2021. The Dynamics of Trust and Verbal Anthropomorphism in Human-Autonomy Teaming. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582655

[13] Nancy J. Cooke, Jamie C. Gorman, Christopher W. Myers, and Jasmine L. Duran. 2013. Interactive Team Cognition - Cooke - 2013 - Cognitive Science - Wiley Online Library. 37, 2 (2013), 255–285. https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12009

[14] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies* 104 (July 2019), 428–442. https://doi.org/10.1016/j.trc.2019.05.025

[15] Connor Esterwood and Lionel P. Robert Jr. 2023. Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142 (May 2023), 107658. https://doi.org/10.1016/j.chb.2023.107658

[16] Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths, and Anca D. Dragan. 2020. Pragmatic-Pedagogic Value Alignment. In *Robotics Research*, Nancy M. Amato, Greg Hager, Shawna Thomas, and Miguel Torres-Torriti (Eds.). Springer International Publishing, Cham, 49–57.

[17] Yaohui Guo, Cong Shi, and Xi Jessie Yang. 2021. Reverse Psychology in Trust-Aware Human-Robot Interaction. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4851–4858. https://doi.org/10.1109/LRA.2021.3067626

[18] Yaohui Guo, X. Yang, and Cong Shi. 2023. Enabling Team of Teams: A Trust Inference and Propagation (TIP) Model in Multi-Human Multi-Robot Teams. In *Robotics: Science and Systems XIX*. Robotics: Science and Systems Foundation. https://doi.org/10.15607/RSS.2023.XIX.003

[19] Yaohui Guo and X. Jessie Yang. 2021. Modeling and Predicting Trust Dynamics in Human-Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* (12 2021). https://doi.org/10.1007/s12369-020-00703-3

[20] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative Inverse Reinforcement Learning. https://doi.org/10.48550/ARXIV.1606.03137

[21] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183.

[22] Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. 2023. "What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/3568162.3576964

[23] G. Klien, D.D. Woods, J.M. Bradshaw, R.R. Hoffman, and P.J. Feltovich. 2004. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems* 19, 6 (Nov. 2004), 91–95. https://doi.org/10.1109/MIS.2004.74 Conference Name: IEEE Intelligent Systems.

[24] Michael Laakasuo, Jussi Palomäki, Anton Kunnari, Sanna Rauhala, Marianna Drosinou, Juho Halonen, Noora Lehtonen, Mika Koverola, Marko Repo, Jukka Sundvall, Aku Visala, and Kathryn B. Francis. 2023. Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology* 53, 1 (2023), 108–128. https://doi.org/10.1002/ejsp.2890 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2890

[25] Huao Li, Tianwei Ni, Siddharth Agrawal, Fan Jia, Suhas Raja, Yikang Gui, Dana Hughes, Michael Lewis, and Katia Sycara. 2021. Individualized Mutual Adaptation in Human-Agent Teams. *IEEE Transactions on Human-Machine Systems* 51, 6 (2021), 706–714. https://doi.org/10.1109/THMS.2021.3107675

[26] Ruikun Luo, Yifan Weng, Yifan Wang, Paramsothy Jayakumar, Mark J. Brudnak, Victor Paul, Vishnu R. Desaraju, Jeffrey L. Stein, Tulga Ersal, and X. Jessie Yang. 2021. A workload adaptive haptic shared control scheme for semi-autonomous driving. *Accident Analysis & Prevention* 152 (2021), 105968. https://doi.org/10.1016/j.aap.2020.105968

[27] Joseph B. Lyons and Svyatoslav Y. Guznov. 2019. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science* 20, 4 (2019), 440–458. https://doi.org/10.1080/1463922X.2018.1491071 arXiv:https://doi.org/10.1080/1463922X.2018.1491071

[28] Joseph B. Lyons, Izz aldin Hamdan, and Thy Q. Vo. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138 (Jan. 2023), 107473. https://doi.org/10.1016/j.chb.2022.107473

[29] Joseph B. Lyons, Thy Vo, Kevin T. Wynne, Sean Mahoney, Chang S. Nam, and Darci Gallimore. 2021. Trusting Autonomous Security Robots: The Role of Reliability and Stated Social Intent. *Human Factors* 63, 4 (2021), 603–618. https://doi.org/10.1177/0018720820901629 arXiv:https://doi.org/10.1177/0018720820901629 PMID: 32027537.

[30] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. 2017. Should Robots be Obedient? http://arxiv.org/abs/1705.09990 arXiv:1705.09990 [cs].

[31] Bonnie Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39 (04 1996), 429–60. https://doi.org/10.1080/00140139608964474

[32] Andrew Y. Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *in Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 663–670.

[33] Charles Pippin and Henrik Christensen. 2014. Trust modeling in multi-robot patrolling. *Proceedings - IEEE International Conference on Robotics and Automation*, 59–66. https://doi.org/10.1109/ICRA.2014.6906590

[34] Lindsay Sanneman and Julie A. Shah. 2023. Validating metrics for reward alignment in human-autonomy teaming. *Computers in Human Behavior* 146 (Sept. 2023), 107809. https://doi.org/10.1016/j.chb.2023.107809

[35] Thomas B. Sheridan. 2016. Human–Robot Interaction: Status and Challenges. *Human Factors* 58, 4 (2016), 525–532. https://doi.org/10.1177/0018720816644364 Publisher: SAGE Publications Inc.

[36] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* 39, 2-3 (2020), 233–249. https://doi.org/10.1177/0278364919866905 arXiv:https://doi.org/10.1177/0278364919866905

[37] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 109–116. https://doi.org/10.1109/HRI.2016.7451741

[38] Ning Wang, David V Pynadath, Susan G Hill, Ning Wang, and David V Pynadath. 2015. Building Trust in a Human-Robot Team with Automatically Generated Explanations. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)* 15315 (2015), 1–12.

[39] Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 221–228.

[40] X. Jessie Yang, Yaohui Guo, and Christopher Schemanske. 2023. From Trust to Trust Dynamics: Combining Empirical and Computational Approaches to Model and Predict Trust Dynamics in Human-Autonomy Interaction. In *Human-Automation Interaction: Transportation*, Vincent G. Duffy, Steven J. Landry, John D.

Lee, and Neville A. Stanton (Eds.). 253–265.
[41] X. Jessie Yang, Christopher Schemanske, and Christine Searle. 2023. Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors* 65, 5 (2023), 862–878. https://doi.org/10.1177/00187208211034716
[42] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. ACM, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230
[43] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science Robotics* 7, 68 (2022), eabm4183. https://doi.org/10.1126/scirobotics.abm4183
[44] Zahra Zahedi, Mudit Verma, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Trust-Aware Planning: Modeling Trust Evolution in Iterated Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Stockholm Sweden, 281–289. https://doi.org/10.1145/3568162.3578628
[45] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. 2010. Modeling Interaction via the Principle of Maximum Causal Entropy. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, Madison, WI, USA, 1255–1262.