Value Alignment and Trust in Human-Robot Interaction: Insights from Simulation and User Study

Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang

Abstract With the advent of AI technologies, humans and robots are increasingly teaming up to perform collaborative tasks. To enable smooth and effective collaboration, the topic of value alignment (operationalized herein as the degree of dynamic goal alignment within a task) between the robot and the human is gaining increasing research attention. Prior literature on value alignment makes an inherent assumption that aligning the values of the robot with that of the human benefits the team. This assumption, however, has not been empirically verified. Moreover, prior literature does not account for human's trust in the robot when analyzing human-robot value alignment. Thus, a research gap needs to be bridged by answering two questions: How does alignment of values affect trust? Is it always beneficial to align the robot's values with that of the human? We present a simulation study and a human-subject study to answer these questions. Results from the simulation study show that alignment of values is important for trust when the overall risk level of the task is high. We also present an adaptive strategy for the robot that uses Inverse Reinforcement Learning (IRL) to match the values of the robot with those of the human during interaction. Our simulations suggest that such an adaptive strategy is able to maintain trust across the full spectrum of human values. We also present results from an empirical study that validate these findings from simulation. Results indicate that real-time personalized value alignment is beneficial to trust and perceived performance by the human when the robot does not have a good *prior* on the human's values.

S. Bhat · X. J. Yang (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: shreyasb@umich.edu; xijyang@umich.edu

J. B. Lyons
Air Force Research Laboratory, Dayton, OH, USA
e-mail: joseph.lyons.6@us.af.mil

C. Shi
University of Miami, Miami, FL, USA
e-mail: congshi@bus.miami.edu

Keywords Human-robot teaming · Value alignment · Trust in HRI

1 Introduction

Robots are progressively becoming more essential in our everyday activities, establishing their roles in numerous fields such as health care, manufacturing, education, and domestic aid, among others. With this ongoing integration, robots are transitioning from being viewed merely as devices executing specific tasks to being recognized as cooperative allies that work alongside humans. Within the context of human-robot teaming, research studying the trust of the human on their robot counterpart is becoming increasingly important [6, 11, 18, 56, 62]. Without an adequate level of trust, we cannot realize the full potential of human-robot teams.

Substantial research efforts have been directed toward creating robots that demonstrate trustworthy behavior and exploring ways to predict and influence a human's trust in robots. For example, one particular line of inquiry is concentrated on generating explanations for robotic actions [21, 41, 42, 58, 59], which often results in increased perception of reliability and, consequently, greater trust in these machines. Other research areas are focused on developing algorithms that can predict trust in real time [24, 26, 57, 61, 64], understanding the dynamics of trust [20, 63, 64] devising strategies to mend broken trust [22, 27, 31], and creating planning methods that take trust into consideration [2, 8, 16, 17, 51, 67].

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 [54, 66]. 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 [12, 19, 28] or demonstrations [4, 23, 28].

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 [5, 38]. 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 [35]. 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 as they negotiate complex social situations [33].

In this chapter, we present results from two studies—a simulation and an empirical verification of the results from simulation. The goal of these studies was to examine the effect of the degree of alignment of values (seen through the lens of reward weights) between the human and the robot in a task with conflicting goals. The results of the simulation study indicate that misalignment of values between the human and the robot result in a loss of trust, but this loss is only apparent under high-

risk scenarios. Under low-risk scenarios, the degree of misalignment is not very influential on trust. In high-risk scenarios, therefore, a robot that adapts its values to that of the human will result in higher trust by the human. We empirically verify this claim through a human-subjects study conducted in a high-risk scenario. The results of this study show that the adaptive-learner strategy presented in our work is able to learn the human's values during interaction and this, build and maintain trust, when compared to a non-adaptive strategy.

2 Related Work

There are two bodies of work closely related to the studies presented in this chapter. One deals with using trust as a decision variable to drive the decision-making processes of robots in human-robot teams and the other deals with ensuring the alignment of values between the human and the robot. A brief review of these fields is presented in this section.

2.1 *Trust-Aware Decision-Making*

In the past few years, significant work has been done to create algorithms for robots that take into account the level of trust humans have in them while performing collaborative tasks [8, 25, 30, 36, 48, 60, 65]. These studies incorporate a representation of a human's trust in the robot into the decision-making processes of the robot and utilize quantitative models of trust and predict human behavior depending on this level of trust.

A quantitative trust model based on the beta distribution is presented in Guo et al. [24]. This model was then used in a simulation study done by Guo et al. [25] to demonstrate the need for using a trust-gaining reward term in the reward function of the robot to encourage trust-gaining behavior. Bhat et al. [8] then conducted a human-subjects experiment using this model and found three types of trust dynamics and associated personal characteristics with the type of trust dynamics exhibited by a person, highlighting the need for robot adaptation given individual differences in trust dynamics.

Trust-aware decision-making has also been made possible by modeling the interaction as a partially observable Markov decision process (POMDP) with trust as the partially observable state variable [1, 2, 17, 65]. Akash et al. [1] present a trust-workload POMDP model that can be learned through interaction data and can be solved to generate optimal policies for the robot that control the level of transparency of the robot's interface [2]. Chen et al. [17] provide a trust-POMDP model that can be solved for getting optimal policies for a robot. In a collaborative pick-and-place task with a robotic arm, they use their model to show trust-gaining behaviors learnt by the robot when it senses that the human's trust is low. On the

other hand, when trust is high, the robot selects risky but high-reward actions to gain the maximum reward.

Zahedi et al. [67] formulate a meta-MDP problem with trust as a decision variable to choose the robot's behavior enabling trust-aware decision-making for the robot. A robot using this framework is shown to display suboptimal but trustworthy behaviors over optimal but untrustworthy behaviors when the human's trust is low. At the same time, if there is enough trust, the robot chooses the optimal action plans to maximize reward gain.

Our work differs from these studies in two main aspects. One, a major part of prior work deals with scenarios in which the reward function is explicitly known by the team. Further, it is assumed that this reward function is shared by the human and the robot. In reality, however, each individual may have their own preferences about the best way of reaching the goal, which are realized as their own reward functions. This "personal" reward function may be different from the robot's "designer-set" reward function. More natural exchanges between the robot and human may not be ones in which the values are known or aligned. In our work, we take into account this possible difference in objectives between the robot and the human and study its effect on the human's trust and team performance. We also provide results when the robot uses an adaptive interaction strategy to match its reward function to that of the human. Secondly, a majority of trust-aware decision-making research deals with robot-worker human-supervisor scenarios in which the human can choose to either observe and interrupt the robot or to focus on their own task. In our work, we are working with a robot-recommender human-decision-maker scenario in which the robot can only provide recommendations to the human on which action to select. The final say in which action to implement lies with the human.

2.2 Value Alignment

Another area of study that has received considerable attention within the HRI community over the past few years deals with the problem of aligning the values/goals of the robot with that of its human counterpart [3, 13, 15, 23, 28, 44, 45, 55, 66].

Bobu et al. [14] present a brief review of literature on aligning human and robot representations and conclude by emphasizing the need for future research in this field and providing some directions for future work. Natarajan et al. [47] study the effects of adapting the driving behavior of autonomous vehicles (AVs) to the preferences of the users. Their results indicate that AVs that continuously adapt their driving style to match the dynamic preferences of users result in higher level of trust by the users. Mechergui and Sreedharan [43] propose and evaluate an interactive learning algorithm that solves the goal alignment problem when the human possesses an incorrect model of the robot's behavior.

The value alignment problem is formulated as a Cooperative Inverse Reinforcement Learning (CIRL) problem in Hadfield-Menell et al [28]. They argue that the traditional method of training an inverse reinforcement learning agent through

human demonstrations in isolation is not an optimal way to teach. They show that solutions to the CIRL problem result in behaviors for the agent that are akin to active teaching and active learning. A solution to the CIRL problem is presented in Fisac et al. [23]. This solution promotes pedagogic behavior by the human teacher, actively taking into account the learning process of the agent. The agent, on the other hand, expects this pedagogic behavior and acts pragmatically on it, resulting in efficient learning of behaviors.

Christiano et al. [19] provide an algorithm that efficiently uses human preferences to learn robot behaviors that reach the goal states in reinforcement learning tasks.

There are two main differences between the work presented in this chapter and prior literature in value alignment. Firstly, as noted earlier, most prior work deals with a human-supervisor robot-worker scenario wherein the robot performs some task while the human observes it and can interrupt the operation if they want (i.e., supervisory control). We study scenarios where the robot recommends actions to the human and the final choice of action to be implemented lies with the human (i.e., decision support). Thus, in order to know when the human will accept the recommendation and when they will not, it is important for the robot to model human decision-making behavior. One way of doing this is through the level of trust of the human on the robot's recommendations. The higher the trust, the higher the likelihood of the human accepting the recommendation and vice versa. Thus, we need an 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. To the best of our knowledge, this is one of the first studies looking into this effect.

3 Simulation Study

As a first step in this research direction, we designed a study in simulation that looked at the effect of the degree of (mis)alignment of reward weights between the human and the robot on the human's trust. This section describes the details of this simulation study and presents our major results

3.1 Human-Robot Teaming Task

We focus on an Intelligence, Surveillance, and Reconnaissance (ISR) mission for the human-robot team. The task involves sequentially searching through M search sites for potential threats. At each site, an *intelligent* robot provides recommendations to the human on whether or not s/he should use an armored robot for protection. In this chapter, the term "robot" refers to the intelligent recommender robot, unless

otherwise specified. Breaching a site without the armored robot is faster as the human does not lose time in deploying the armored robot. At the same time, it is risky as the human may be harmed if a threat is encountered without protection with the armored robot. Using the armored robot is safer as it protects the human from the threat. However, it is also slower to use the armored robot as it takes time to deploy it. The objective of the team is to minimize damage to the human while also minimizing mission completion time.

3.2 Trust-Aware Markov Decision Process

We model the interaction between the human and the robot as they sequentially search through the town as a Trust-Aware Markov Decision Process (trust-aware MDP). It consists of states, actions, reward function, transition function, and a human trust-behavior model.

3.2.1 States

We define the trust of the human in the robot as the state. To quantify trust, we use the model from [8] which defines the trust t_i at site i to follow a Beta distribution with parameters α_i and β_i, i.e.,

$$t_i \sim \text{Beta}(\alpha_i, \beta_i). \tag{1}$$

3.2.2 Actions

At any site i, the robot has two actions available: Recommend to USE the armored robot or recommend to NOT USE the armored robot. So, we say that, $a^r \in \{0, 1\}$. Here, action 0 corresponds to recommending to not use the armored robot and action 1 to recommending to use the armored robot.

3.2.3 Rewards

We define the rewards for each agent as a weighted sum of the health loss reward and time loss reward. The reward at site i for agent o is given by,

$$R_i^o\left(a_i^h, D_i\right) = -w_h^o h\left(a_i^h, D_i\right) - w_c^o c\left(a_i^h\right). \tag{2}$$

Here, the weights w_h^o and w_c^o are the weights associated with the health loss reward and the time loss reward, respectively, for agent $o \in \{h, r\}$. The superscript r shows

that these values are for the robot, and h shows that the values are for the human. a_i^h is the action chosen by the human at site i. D_i is also a random variable which represents the presence of threat inside site i. $h(\cdot, \cdot)$ and $c(\cdot)$ are functions that give the health loss cost and the time loss cost, respectively. In our simulation, we defined these functions as follows, $h(1, 1) = 0, h(0, 1) = 10, h(1, 0) = 0, h(0, 0) = 0$ and $c(0) = 0, c(1) = 10$.

3.2.4 Transition Model

The transition model describes the evolution of trust during the interaction. We use the reward-based performance metric from [8] to define state transitions.

$$\alpha_i = \alpha_{i-1} + P_i v^s, \tag{3}$$

$$\beta_i = \beta_{i-1} + (1 - P_i)v^f. \tag{4}$$

With the performance P_i defined as,

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

3.2.5 Human Trust-Behavior Model

The human behavior model encodes how a human agent responds to recommendations made by the robot. We state that the probability of the human to accept the recommendation given by the robotic agent is directly proportional to their level of trust. If the human does not accept the recommendation, s/he chooses the action that gives a higher immediate expected reward with a higher probability. More precisely, let a_i^r and a_i^h denote the action recommended by the robot and the action chosen by the human at site i, respectively. Then,

$$\mathbb{P}\left(a_i^h = a | a_i^r = a\right) = t_i + (1 - t_i)p_i^a,$$

$$\mathbb{P}\left(a_i^h = 1 - a | a_i^r = a\right) = (1 - t_i)\left(1 - p_i^a\right). \tag{6}$$

Here t_i is the trust level at the ith site and p_i^a is given by,

$$p_i^a = \frac{\exp(\kappa E[R_i^h(a)])}{\exp(\kappa E[R_i^h(a)]) + \exp(\kappa E[R_i^h(1 - a)])}. \tag{7}$$

Here, the superscript a can be either 0 or 1 corresponding to the two actions. κ controls the effect of the rewards on the probability p_i^a. In human behavior modeling

literature, this is often referred to as the rationality coefficient [7, 23, 66]. The higher the value of κ, the more likely it is that the human will choose the action with the higher reward. On the other hand, $\kappa = 0$ represents a uniformly random human, one who chooses each action with equal probability. We call the resulting model the *bounded rationality disuse* model, since it combines concepts from both models.

3.2.6 Value Iteration

The robot solves the trust-aware MDP via value iteration. It selects the action by maximizing the expected reward at the current step summed together with a discounted value of the next state at the next stage.

$$Q_i(s_i, a) = E\left[R_i^r(a)\right] + \sum_{s_{i+1} \in S} \gamma \mathbb{P}(s_{i+1}|s_i, a) V_{i+1}(s_{i+1}). \tag{8}$$

$$V_i(s_i) = \max_a Q_i(s_i, a). \tag{9}$$

At the final stage, the action that gives the maximum immediate expected reward is chosen.

$$V_N(s_N) = \max_a E\left[R_N^r(a)\right]. \tag{10}$$

3.2.7 Bayesian Inverse Reinforcement Learning

The robot uses a Bayesian inverse reinforcement learning [53] framework with the assumed human trust-behavior model to learn the human's preferences between saving health and saving time. The robot maintains a belief $b_i(w_h^h)$ over the possible health reward weights of the human. Here, the subscript i denotes that this is the robot's belief just before the ith site is searched. After observing the action selected by the human, the robot updates this belief using Bayes rule.

If at site i, we observe that the human chooses the recommended action, we update the belief according to

$$b_{i+1}(w) \propto \mathbb{P}\left(a_i^h = a|w_h^h = w, a_i^r = a, t_i\right) b_i(w)$$

$$\propto \left(t_i + (1 - t_i)p_i^a\right) b_i(w). \tag{11}$$

If, at site i, we observe that the human chooses the action opposite to the one recommended, we update the belief according to

$$b_{i+1}(w) \propto \mathbb{P}\left(a_i^h = 1 - a|w_h^h = w, a_i^r = a, t_i\right) b_i(w)$$

$$\propto (1 - t_i)(1 - p_i^a)b_i(w). \tag{12}$$

Here, p_i^a is computed according to Eq. 7, substituting weight w in the expected reward computation. We use the mean of the belief distribution as a proxy for the human's health weight. We always set $w_c^h = 1 - w_h^h$.

3.2.8   Threats and Threat Levels

We follow the following strategy to set the threat levels at the search sites. First, a prior probability of threat presence d is fixed across all sites. Then, at each site independently, threat presence is determined by a Bernoulli distribution with d as the parameter. The drone gets an updated threat level after scanning a site. This updated threat level is generated by a beta distribution with a peak at 0.1 for when a threat is not present and a peak at 0.9 for when a threat is present inside the site.

3.2.9   Simulating Human Decision and Feedback

We sample from the set of trust parameters Θ that was generated from a previous study [8] to get $\theta = (\alpha_0, \beta_0, v^s, v^f)$ for the simulated human. These parameters are unknown to the robot, and it estimates them via maximum likelihood estimation (for details, refer to [8]). The human also has reward weights w_h^h, $w_c^h (:= 1 - w_h^h)$ which are unknown to the robot. The human gives trust feedback by sampling from the beta distribution trust update model, using the observed rewards associated with the recommendation to judge the performance of the robot. The human chooses his/her action using the bounded rationality disuse model with $\kappa = 1$.

3.3   Results: Simulation Study

This section provides details about the major results from this simulation study.

3.3.1   Regions in End-of-Mission Trust

We observed four distinct regions in the end-of-mission trust as a function of the reward weights of the human and the robot (Fig. 1). If a risk-averse human is paired with a risk-taking robot, the end-of-mission trust can get very low (the region in the top left of the plots in Fig. 1) and vice versa. This is especially true when the risk level is higher, as can be seen in the higher difference in trust levels when the values align and when they don't in Fig. 1b compared to the same in Fig. 1a. In the figure on the left, the probability of threat being present in any site was 0.3, while the same was 0.7 in the figure on the right.
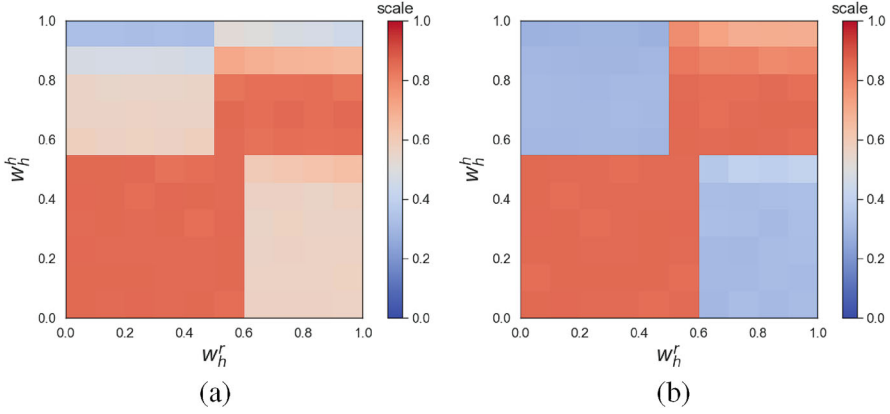
Fig. 1 The observed regions in end-of-mission trust as a function of the health weights of the human w_h^h and the robot w_h^r. The figure on the left is the simulation result when there is a relatively low chance of threat presence at any search site ($d = 0.3$). The figure on the right is when there is a higher chance of threat presence at any search site ($d = 0.7$). (**a**) $d = 0.3$. (**b**) $d = 0.7$

3.3.2   Effect of Threat Level

Since the results from the last section show some potential effects of the level of threat associated with the mission on the end-of-mission trust, we ran another set of simulations, this time, fixing the reward weights of the human and the robot (i.e., the trust region) and varying the level of threat. Figure 2 shows plots of the end-of-mission trust as a function of the prior threat level d for the four regions of trust. Further, it is seen that if the values are roughly aligned, trust increases. If the values are not aligned, trust is still high when the overall risk is low (at low threat levels d). However, the human starts to lose trust in the robot if the values are misaligned and the risk level is high.

Figure 3 shows an example when both the human and the robot are extremely risk-averse. In this case, although there is alignment in the reward weights of the human and the robot, the level of trust is generally very low. This is because the robot tries to maximize expected reward, but the human is judging performance through observed rewards. At lower risk levels, it would be better to be cautious in the long term, but it is more likely to get better rewards in the short term by being risky. Thus, it seems to the human that the robot is overly cautious, resulting in a decrease in trust.

3.3.3   Effect of Adapting to the Human's Rewards

We can update the reward weights of the robot at each time step by using the distribution on the reward weights of the human maintained by the robot (see Sect. 3.2.7). This results in an "adaptive" interaction strategy in which the robot
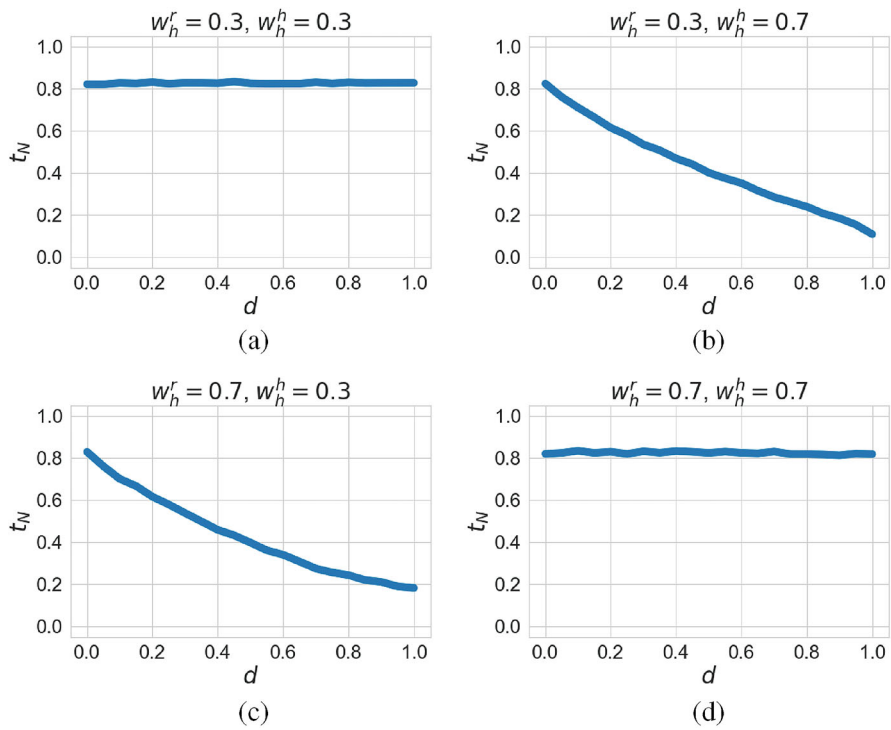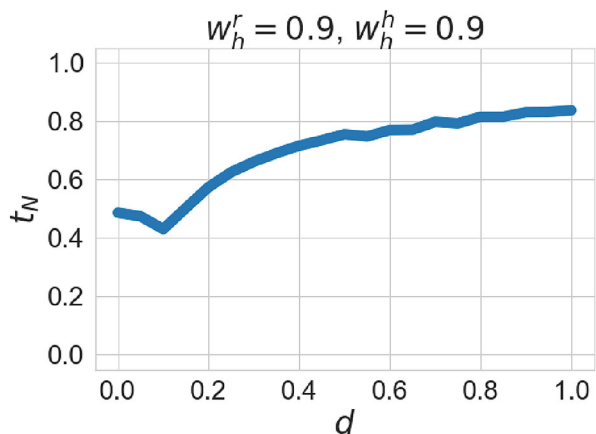
Fig. 2 The effect of the prior probability of threat presence in any house d on the end-of-mission trust, after fixing a trust region. (**a**) Region 1. (**b**) Region 2. (**c**) Region 3. (**d**) Region 4

Fig. 3 The effect of prior probability of threat presence in any house d on the end-of-mission trust when the human and the robot are both extremely risk-averse
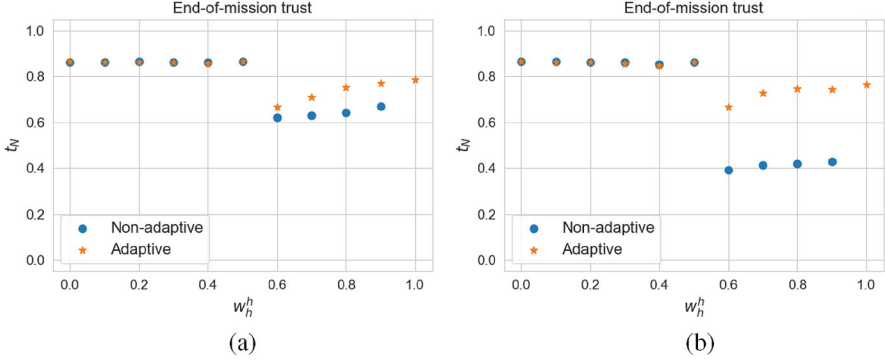
Fig. 4 Comparing the adaptive strategy with the non-adaptive strategy in the end-of-mission trust feedback given by the simulated human for two different levels of threat. The non-adaptive strategy sets the robot's health reward weight to 0.5. (**a**) $d = 0.3$. (**b**) $d = 0.7$

adapts to human values throughout their interaction. We ran simulations when the robot followed such a strategy, starting with a uniform distribution on the reward weights. We can then compare this adaptive strategy with a non-adaptive strategy of the robot with $w_h^r = 0.5$ (the mean of the initial uniform distribution) to see the effect of adapting to human preferences.

Figure 4 shows the effect of the human's health reward weight on the end-of-mission trust reported by the human at two threat levels: low (0.3) and high (0.7). As can be seen, below $w_h^h < 0.5$, there is no clear difference between trust on the non-adaptive strategy and that on the adaptive strategy. This is because at $w_h^r = 0.5$, the robot has no preference between losing health and losing time. However, in expectation, not using the RARV is always better than using the RARV. Thus, the recommendation matches the human's preferred action (since the human also does not care about losing health when $w_h^h < 0.5$). As a result, trust is high for both non-adaptive and adaptive strategies. On the other hand, however, when $w_h^h > 0.5$, the human prefers to save health over saving time. Thus, the risky recommendations of the non-adaptive strategy reduce trust, while the adaptive strategy is able to learn this preference and is able to maintain the human's trust.

Figure 5 compares the two strategies for two different human preferences (preferring to lose health $w_h^h = 0.3$ and preferring to lose time $w_h^h = 0.7$) as the level of threat is changed. Again, we see similar behavior that when the human prefers to lose health, both strategies result in the same level of trust of the human on the robot. On the other hand, when the human prefers to lose time, the non-adaptive strategy's risky recommendations lead to a loss in trust as the level of threat increases. Again, the adaptive strategy is able to learn this preferences resulting in a much higher level of the human's trust, irrespective of the level of threat.
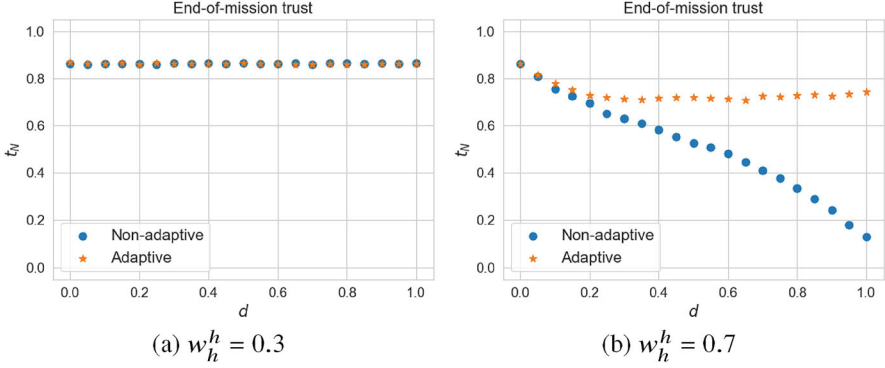
Fig. 5 Comparing the adaptive strategy with the non-adaptive strategy in the end-of-mission trust feedback given by the simulated human for two different levels of threat. The non-adaptive strategy sets the robot's health reward weight to 0.5. (**a**) $w_h^h = 0.3$. (**b**) $w_h^h = 0.7$

3.4 Discussion: Simulation Study

A key takeaway from our results is that, under the bounded rationality disuse model, value alignment is, in general, good for trust. However, at low-risk settings, value alignment may not be as important for gaining trust. This conclusion is somewhat in line with the finding of [40], wherein the authors found increased reliance on automation in high-risk situations compared to low-risk situations. Parkhe and Stewart [50] claim that trust is only necessary when the losses can exceed potential gains (i.e., when the risk is high). We take it a step further by saying that value alignment is only necessary for trust when the risks are high. Further, since humans are generally more concerned with observed outcomes rather than expected rewards, the boundaries of the regions can result in lower end-of-mission trust despite having values aligned between the human and the robot.

In light of our results, we claim that completely aligning the robot's values to the human's may not always be the best, especially when the human is extremely risk-averse. This may result in loss of performance *and* loss of trust. On the other hand, not aligning the values at all will certainly lead to a loss in trust, especially in high-risk situations. Thus, we postulate that there may be merit in looking for a middle ground that trades off between fully aligning the values and not aligning them at all. This could be a very interesting direction for future research.

4 Empirical Study

This section provides details on the human-subjects experiment carried out to empirically validate the results of the simulation study. An extension of this study

Fig. 6 The recommendation interface showing the information about the threat level and the recommendation by the intelligent agent. Also shown, 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. (**a**) The recommendation interface. (**b**) Trust feedback slider

is presented in [9, 10]. Interested readers are encouraged to refer to these studies for further details.

4.1 Testbed

We developed a 3D testbed using the Unreal Engine game development platform. Within the testbed, a human is paired with an autonomous drone and an armored robot to complete the reconnaissance mission described in Sect. 3.1. Figure 6a shows a screenshot of the recommendation interface used by the intelligent agent. It displays the threat level indicated by the drone, the average time to search a site with and without the Robotic Armored Rescue Vehicle (RARV), and the system's recommendation. The participant is then free to select an action by pressing the associated key on the keyboard. The participants are, at all times, displayed the time left to complete the mission and the current health level of the human. They are also displayed the index of the current search site. They are given a total of 25 minutes to complete searching a total of 40 sites. Each time the human encounters a threat without protection from the RARV, they lose 5 points of health. Deploying the RARV takes approximately 15 seconds. After exiting a search site, the participants are asked to report their trust on the recommendation system on a slider (shown in Fig. 6b). They are shown the drone's assessed level of threat, the recommendation given by the system, the action selected by the participant, the ground truth of threat presence, and the time it took them to search the site on the feedback dialog to aid with recall. The mission timer is stopped when the participants are shown the feedback slider. This is to let the participants take their time to assess and report their trust accurately. The threats and threat levels the drone gets after scanning a site are randomly generated while ensuring that there are threats in 23 out of the 40 sites, corresponding to a prior threat level $d = 0.575$.

4.2 Conditions

We designed three interaction strategies for the intelligent agent:

- **Non-learner:** The intelligent agent does not learn the reward weights of the human. It assumes that the human shares the same reward weights as itself 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 during interaction with each human. 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 then adopts this learnt reward weights as its own.

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. They were not informed which interaction strategy the robot was using. They were only informed of their dual goal of minimizing damage to the soldier while also completing the mission as quickly as possible. A 3×3 Latin Square design was used to minimize any learning effects.

4.3 Participants

We collected data from a total of 24 participants (Age: Mean 21.4 years, *SD* 2.3, 12 Female). All participants were students from the University of Michigan.

4.4 Measures

4.4.1 Pre-experiment Measures

Prior to the experiment, participants completed 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 type of 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. 6b for the exact question asked). The slider values were between 0 and 100 with a step size of 2 points. Additionally, for every trial, we measured whether the participants agreed with the intelligent agent's recommendation. With the trial-based data, we measured the following:

- **Average Trust:** This was calculated as the empirical mean trust $\frac{1}{M} \sum_{i=1}^{M} t_i$.
- **Number of Agreements:** 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 completed survey assessing the following items.

- **Post-mission trust questionnaire:** Measured using Muir's trust questionnaire [46]. It has nine questions, each with a slider with a range between 0 and 100.
- **Post-mission Reliance Intentions:** Measured using the scale developed by Lyons and Guznov [39]. We used six of the ten items that were relevant for this task. Each item was rated on a 7-point Likert scale.
- **Workload:** Workload was measured using the NASA Task Load Index (NASA TLX) scale [29, 37, 38]. We used five of the six 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. The weights used were the participants rated preferences at the beginning of the mission.

$$\text{Performance} = \hat{w}_h^h \cdot (\%h_M) + \hat{w}_c^h \cdot (100 - \%c_M). \tag{13}$$

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.

Table 1 Mean $\pm$ standard deviation (SD) of measures for the interaction strategies

	Non-learner	Non-adaptive learner	Adaptive-learner
Average trust $\frac{1}{M}\sum_{i=1}^{M} t_i$ [a]	0.42 ± 0.20	0.45 ± 0.22	0.65 ± 0.20
End-of-mission trust t_M [a]	0.33 ± 0.30	0.35 ± 0.29	0.64 ± 0.30
Muir's trust questionnaire[a]	34.51 ± 19.79	34.97 ± 17.41	60.57 ± 23.71
Number of agreements[a]	28.17 ± 5.49	28.67 ± 4.65	35.62 ± 2.93
Reliance intentions scale[a]	2.16 ± 1.22	2.38 ± 1.16	3.58 ± 1.32
Workload[a]	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

[a] $p < 0.05$

4.5 Results: Empirical Study

This section summarizes our results and discusses the implications (see Table 1 for an overview). Repeated measure analyses of variance (ANOVAs) were conducted to compare the three interaction strategies. Greenhouse-Geisser corrections to the degrees of freedom were made for the measures that failed Mauchly's test of sphericity.

4.6 Trust, Agreement, and Reliance Intention

Figure 7a, b show the comparisons of the three strategies in trust. Repeated measures ANOVA revealed significant differences between the three strategies in average trust ($F(2, 46) = 14.161$, $p < 0.001$) 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 and post-mission trust compared to the non-learner strategy ($p < 0.001$ and $p < 0.001$, respectively) and compared to the non-adaptive learner strategy ($p = 0.003$ and $p < 0.001$, respectively).

Regarding the number of agreements (Fig. 7c), 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 ($p < 0.001$).

Comparing reliance intentions (Fig. 7d), 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 ($p = 0.004$).
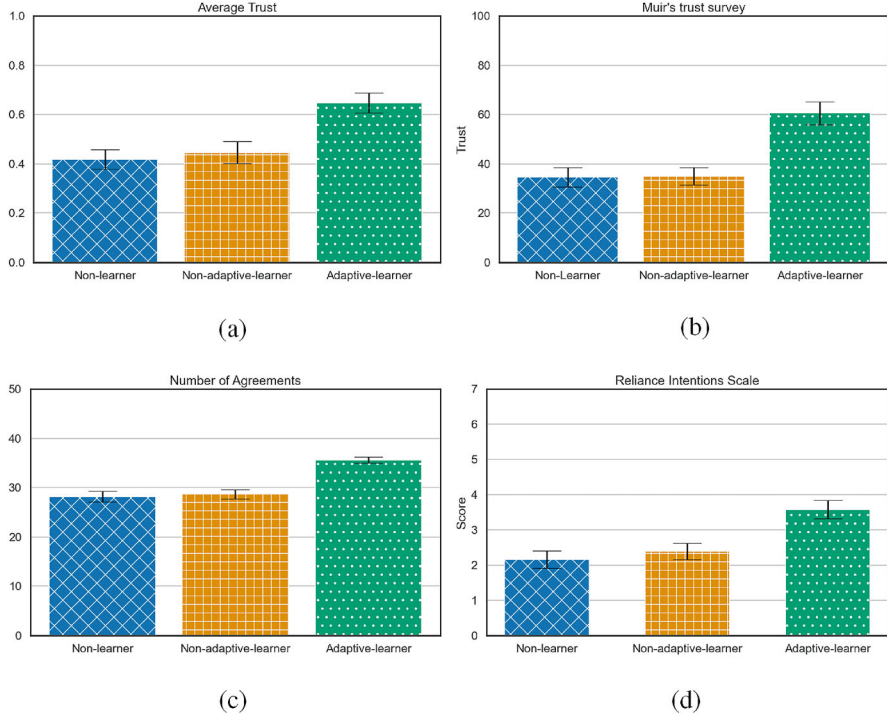
Fig. 7 Comparing the human's trust and reliance intentions on the robot when the robot follows different interaction strategies. (**a**) Average trust $\frac{1}{M}\sum_{i=1}^{M} t_i$. (**b**) Subjective Trust—Muir's scale. (**c**) Number of agreements. (**d**) Reliance intentions

4.7 Performance

There seemed to be an upward trend from non-learner to non-adaptive learner and to adaptive learner. Unfortunately, it did not reach significance.

4.8 Workload

Comparing the average workload (1) across the three interaction strategies showed a significant difference ($F(2, 46) = 10872$, $p < 0.001$). Figure 8 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
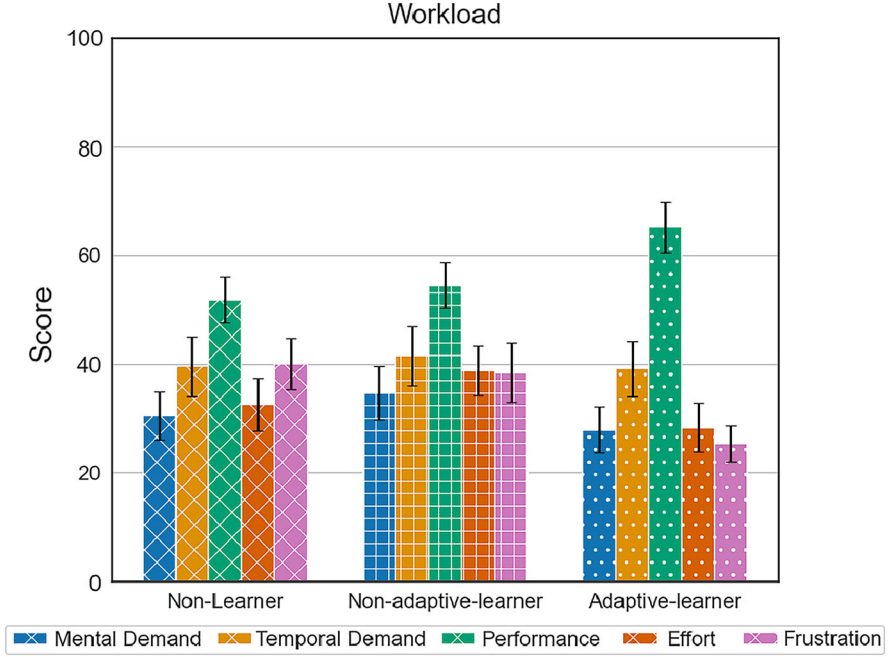
Fig. 8 Exp 2: Responses on the NASA TLX scale

frustration compared to the non-adaptive learner ($p = 0.032$) and to the non-learner ($p = 0.017$) strategies.

5 Discussion: Empirical Study

In this study, we empirically verified the results of the simulation study by having participants interact with non-adaptive strategies which was not value aligned with them and an adaptive strategy that continuously adapts to the human's values during interaction. Results from this study correspond to the case with a high level of risk $d = 0.575$ from the simulation study in which we saw that adapting to human values kept trust high while trust is lost while interacting with a misaligned robot.

Additionally, we also observed that human workload is lower when interacting with the adaptive-learner strategy compared to that when interacting with the non-learner and non-adaptive learner strategies. Although the objective performance was the same across the three strategies, the participants perceived that they were performing better with the adaptive learner strategies. The similar objective performance across the three strategies could be attributed to the information provided to the participants, which was usually enough to make their own decisions, regardless

of the robot's recommendations. Thus, the human action selection policies were very similar across the three strategies. Only the adaptive learner strategy was able to learn and match this human policy resulting in a higher number of agreements with the robot's recommendations, which we postulate, led to lower workload and higher trust.

6   Conclusion

Most prior work in value alignment in HRI [23, 28, 43, 66] makes one key assumption: Aligning the robot's values to that of the human's is beneficial. This may be true when we look at the human in isolation; the human's rewards would increase in this case. However, when looking at it from a teaming perspective, we do not know if the human's personal values are the best for handling the task at hand [3, 32, 44]. Thus, this assumption needs to be checked in the context of human-robot teams. Our study is one of the first to attempt to study this assumption. Our simulation results indicate that value alignment is beneficial for trust only when there is a high level of risk in the task involved. When the level of risk is low, even value misaligned robots can lead to high trust. Our empirical study shows that value alignment under high risk is indeed beneficial for trust and workload.

We also explore the idea of incorporating a layer of trust within the value alignment problem. We theorize that since there is a tradeoff between degree of value alignment and trust and since it has been shown that trust is a predictor of behavior in humans [34, 49], it is important for the robot to try and find a balance between aligning itself to the human's values and maintaining trust. We hope to see more work done toward this goal in other domains like shared control, social robotics, rehabilitation robotics, etc.

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 task consists of binary actions: using or not using an armored robot. 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. There could be effects like satisficing [52] and heuristics which would make defining the performance metric 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 tasks where the two objectives are more balanced.

Finally, our scenario, which entails a tradeoff 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 and Shield AI's NOVA 2 program involving the use of small drones for surveillance and decision support. 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.

References

1. K. Akash, K. Polson, T. Reid, N. Jain, Improving human-machine collaboration through transparency-based feedback – part I: human trust and workload model. IFAC-PapersOnLine **51**(34), 315–321 (2019). 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018
2. K. Akash, T. Reid, N. Jain, Improving human-machine collaboration through transparency-based feedback – part II: control design and synthesis. IFAC-PapersOnLine **51**(34), 322–328 (2019). 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018
3. T. Arnold, D. Kasenberg, M. Scheutz, Value alignment or misalignment - what will keep systems accountable?, in *AAAI Workshops* (2017). https://api.semanticscholar.org/CorpusID:1725783
4. S. Arora, P. Doshi, A survey of inverse reinforcement learning: challenges, methods and progress. Artif. Intell. **297**, 103500 (2021). https://doi.org/10.1016/j.artint.2021.103500, https://linkinghub.elsevier.com/retrieve/pii/S0004370221000515
5. H. Azevedo-Sa, S.K. Jayaraman, X.J. Yang, L.P. Robert, D.M. Tilbury, Context-adaptive management of drivers' trust in automated vehicles. IEEE Rob. Autom. Lett. **5**(4), 6908–6915 (2020). https://doi.org/10.1109/LRA.2020.3025736
6. H. Azevedo-Sa, X.J. Yang, L.P. Robert, D.M. Tilbury, A unified bi-directional model for natural and artificial trust in human–robot collaboration. IEEE Rob. Autom. Lett. **6**(3), 5913–5920 (2021). https://doi.org/10.1109/LRA.2021.3088082
7. C.L. Baker, J.B. Tenenbaum, Chapter 7 - modeling human plan recognition using bayesian theory of mind, in *Plan, Activity, and Intent Recognition*, ed. by G. Sukthankar, C. Geib, H.H. Bui, D.V. Pynadath, R.P. Goldman (Morgan Kaufmann, Boston, 2014). pp. 177–204. https://doi.org/10.1016/B978-0-12-398532-3.00007-5
8. S. Bhat, J.B. Lyons, C. Shi, X.J. Yang, Clustering trust dynamics in a human-robot sequential decision-making task. IEEE Rob. Autom. Lett. **7**(4), 8815–8822 (2022). https://doi.org/10.1109/LRA.2022.3188902

9. S. Bhat, J.B. Lyons, C. Shi, X.J. Yang, Effect of adapting to human preferences on trust in human-robot teaming. Proc. AAAI Symp. Ser. **2**(1), 5–10 (2024). https://doi.org/10.1609/aaaiss.v2i1.27642, https://ojs.aaai.org/index.php/AAAI-SS/article/view/27642

10. S. Bhat, J.B. Lyons, C. Shi, X.J. Yang, 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* (ACM, Boulder, 2024), pp. 32–41, https://doi.org/10.1145/3610977.3634921, https://dl.acm.org/doi/10.1145/3610977.3634921

11. D.R. Billings, K.E. Schaefer, J.Y.C. Chen, P.A. Hancock, Human-robot interaction: developing trust in robots, in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2012), pp. 109–110. https://doi.org/10.1145/2157689.2157709

12. E. Bıyık, D. Sadigh, Batch active preference-based learning of reward functions (2018). 1810.04303

13. E. Biyik, Learning from humans for adaptive interaction, in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, Sapporo, 2022), pp. 1152–1154. https://doi.org/10.1109/HRI53351.2022.9889436, https://ieeexplore.ieee.org/document/9889436/

14. A. Bobu, A. Peng, P. Agrawal, J.A. Shah, A.D. Dragan, Aligning human and robot representations, in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, Boulder, 2024), pp. 42–54. https://doi.org/10.1145/3610977.3634987, https://dl.acm.org/doi/10.1145/3610977.3634987

15. P. Butlin, AI alignment and human reward, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (ACM, New York, 2021), pp. 437–445. https://doi.org/10.1145/3461702.3462570, https://dl.acm.org/doi/10.1145/3461702.3462570

16. M. Chen, S. Nikolaidis, H. Soh, D. Hsu, S. Srinivasa, Planning with trust for human-robot collaboration, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery*, New York, NY, HRI '18 (2018), pp. 307–315. https://doi.org/10.1145/3171221.3171264

17. M. Chen, S. Nikolaidis, H. Soh, D. Hsu, S. Srinivasa, Trust-aware decision making for human-robot collaboration: model learning and planning. J. Hum.-Rob. Interact. **9**(2) (2020). https://doi.org/10.1145/3359616

18. E.K. Chiou, J.D. Lee, Trusting automation: designing for responsivity and resilience. Hum. Fact. **65**(1), 137–165 (2023). https://doi.org/10.1177/00187208211009995, pMID: 33906505

19. P. Christiano, J. Leike, T.B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences (2023). 1706.03741

20. M.C. Cohen, M. Demir, E.K. Chiou, N.J. Cooke, The dynamics of trust and verbal anthropomorphism in human-autonomy teaming, in *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)* (2021), pp. 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582655

21. N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A.K. Pradhan, X.J. Yang, L.P. Robert, Look who's talking now: implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. Transport. Res. Part C: Emerg. Technol. **104**, 428–442 (2019). https://doi.org/10.1016/j.trc.2019.05.025, http://www.sciencedirect.com/science/article/pii/S0968090X18313640

22. C. Esterwood, L.P. Robert Jr, Three Strikes and you are out!: the impacts of multiple human–robot trust violations and repairs on robot trustworthiness. Comput. Hum. Behav. **142**, 107658 (2023). https://doi.org/10.1016/j.chb.2023.107658, https://www.sciencedirect.com/science/article/pii/S0747563223000092

23. J.F. Fisac, M.A. Gates, J.B. Hamrick, C. Liu, D. Hadfield-Menell, M. Palaniappan, D. Malik, S.S. Sastry, T.L. Griffiths, A.D. Dragan, Pragmatic-pedagogic value alignment, in *Robotics Research*, ed. by N.M. Amato, G. Hager, S. Thomas, M. Torres-Torriti (Springer International Publishing, Cham, 2020), pp. 49–57

24. Y. Guo, X.J. Yang, Modeling and predicting trust dynamics in human-robot teaming: a bayesian inference approach. Int. J. Soc. Rob. (2021). https://doi.org/10.1007/s12369-020-00703-3

25. Y. Guo, C. Shi, X.J. Yang, Reverse psychology in trust-aware human-robot interaction. IEEE Rob. Autom. Lett. **6**(3), 4851–4858 (2021). https://doi.org/10.1109/LRA.2021.3067626

26. Y. Guo, X. Yang, C. Shi, 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* (2023). https://doi.org/10.15607/RSS.2023.XIX.003, http://www.roboticsproceedings.org/rss19/p003.pdf

27. Y. Guo, X.J. Yang, C. Shi, Reward shaping for building trustworthy robots in sequential human-robot interaction, in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2023), pp. 7999–8005. https://doi.org/10.1109/IROS55552.2023.10341904

28. D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell, Cooperative inverse reinforcement learning (2016). https://doi.org/10.48550/ARXIV.1606.03137, https://arxiv.org/abs/1606.03137

29. S.G. Hart, L.E. Staveland, Development of NASA-TLX (task load index): results of empirical and theoretical research, in *Human Mental Workload, Advances in Psychology*, ed. by P.A. Hancock, N. Meshkati, vol. 52 (North-Holland, Amsterdam, 1988), pp. 139–183

30. A. Hussein, S. Elsawah, H.A. Abbass, Towards trust-aware human-automation interaction: an overview of the potential of computational trust models, in *HICSS* (2020), pp. 1–10

31. U.B. Karli, S. Cao, C.M. Huang, "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, 2023), pp. 271–280. https://doi.org/10.1145/3568162.3576964, https://dl.acm.org/doi/10.1145/3568162.3576964

32. M. Kwon, E. Biyik, A. Talati, K. Bhasin, D.P. Losey, D. Sadigh, When humans aren't optimal: robots that collaborate with risk-aware humans, in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, Cambridge, 2020), pp. 43–52. https://doi.org/10.1145/3319502.3374832, https://dl.acm.org/doi/10.1145/3319502.3374832

33. M. Laakasuo, J. Palomäki, A. Kunnari, S. Rauhala, M. Drosinou, J. Halonen, N. Lehtonen, M. Koverola, M. Repo, J. Sundvall, A. Visala, K.B. Francis, Moral psychology of nursing robots: exploring the role of robots in dilemmas of patient autonomy. Eur. J. Soc. Psychol. **53**(1), 108–128 (2023). https://doi.org/10.1002/ejsp.2890, https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2890, https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2890

34. J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance. Hum. Fact. **46**(1), 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50_30392, pMID: 15151155

35. H. Li, T. Ni, S. Agrawal, F. Jia, S. Raja, Y. Gui, D. Hughes, M. Lewis, K. Sycara, Individualized mutual adaptation in human-agent teams. IEEE Trans. Hum.-Mach. Syst. **51**(6), 706–714 (2021). https://doi.org/10.1109/THMS.2021.3107675

36. Y. Li, J. Xu, D. Guo, H. Liu, Trust-aware human–robot fusion decision-making for emergency indoor patrolling. IEEE Trans. Autom. Sci. Eng. (2024), pp. 1–0. https://doi.org/10.1109/TASE.2024.3350639, https://ieeexplore.ieee.org/document/10397549/

37. S. Lu, M.Y. Zhang, T. Ersal, X.J. Yang, Workload management in teleoperation of unmanned ground vehicles: effects of a delay compensation aid on human operators' workload and teleoperation performance. Int. J. Hum.-Comput. Interact. **35**(19), 1820–1830 (2019). https://doi.org/10.1080/10447318.2019.1574059

38. R. Luo, Y. Weng, Y. Wang, P. Jayakumar, M.J. Brudnak, V. Paul, V.R. Desaraju, J.L. Stein, T. Ersal, X.J. Yang, A workload adaptive haptic shared control scheme for semi-autonomous driving. Accid. Anal. Prev. **152**, 105968 (2021). https://doi.org/10.1016/j.aap.2020.105968, https://www.sciencedirect.com/science/article/pii/S0001457520317887

39. J.B. Lyons, S.Y. Guznov, Individual differences in human–machine trust: a multi-study look at the perfect automation schema. Theor. Issues Ergon. Sci. **20**(4), 440–458 (2019). https://doi.org/10.1080/1463922X.2018.1491071

40. J.B. Lyons, C.K. Stokes, Human-human reliance in the context of automation. Hum. Factors **54**(1), 112–121 (2012)

41. J.B. Lyons, T. Vo, K.T. Wynne, S. Mahoney, C.S. Nam, D. Gallimore, Trusting autonomous security robots: the role of reliability and stated social intent. Hum. Factors **63**(4), 603–618 (2021). https://doi.org/10.1177/0018720820901629, pMID: 32027537

42. J.B. Lyons, I.a. Hamdan, T.Q. Vo, Explanations and trust: what happens to trust when a robot partner does something unexpected? Comput. Hum. Behav. **138**, 107473 (2023). https://doi.org/10.1016/j.chb.2022.107473, https://www.sciencedirect.com/science/article/pii/S074756322200293X

43. M. Mechergui, S. Sreedharan, Goal alignment: re-analyzing value alignment problems using human-aware AI. Proc. AAAI Conf. Artif. Intell. **38**(9), 10110–10118 (2024). https://doi.org/10.1609/aaai.v38i9.28875, https://ojs.aaai.org/index.php/AAAI/article/view/28875

44. S. Milli, D. Hadfield-Menell, A. Dragan, S. Russell, Should Robots be Obedient? (2017). http://arxiv.org/abs/1705.09990, arXiv:1705.09990 [cs]

45. N. Moorman, E. Hedlund-Botti, M. Schrum, M. Natarajan, M.C. Gombolay, Impacts of robot learning on user attitude and behavior, in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, Stockholm, 2023), pp. 534–543. https://doi.org/10.1145/3568162.3576996, https://dl.acm.org/doi/10.1145/3568162.3576996

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

47. M. Natarajan, K. Akash, T. Misu, Toward adaptive driving styles for automated driving with users' trust and preferences, in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, Sapporo, 2022), pp. 940–944. https://doi.org/10.1109/HRI53351.2022.9889313, https://ieeexplore.ieee.org/document/9889313/

48. Y. Pang, R. Liu, Trust-aware emergency response for a resilient human-swarm cooperative system, in *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)* (IEEE, New York City, 2021), pp. 15–20. https://doi.org/10.1109/SSRR53300.2021.9597682, https://ieeexplore.ieee.org/document/9597682/

49. R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse. Hum. Factors **39**(2), 230–253 (1997). https://doi.org/10.1518/001872097778543886

50. A. Parkhe, S. Miller, The structure of optimal trust: a comment and some extensions. Acad. Manage. Rev. **25**, 10–11 (2000)

51. C. Pippin, H. Christensen, *Trust Modeling in Multi-robot Patrolling* (IEEE, Piscataway, 2014), pp. 59–66. https://doi.org/10.1109/ICRA.2014.6906590

52. R. Radner, Satisficing, in *Optimization Techniques IFIP Technical Conference Novosibirsk*, July 1–7, 1974 (Springer, Berlin, 1975), pp. 252–263

53. D. Ramachandran, E. Amir, Bayesian inverse reinforcement learning, in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07 (Morgan Kaufmann Publishers Inc., San Francisco, 2007), pp. 2586–2591

54. L. Sanneman, J.A. Shah, Validating metrics for reward alignment in human-autonomy teaming. Comput. Hum. Behav. **146**, 107809 (2023). https://doi.org/10.1016/j.chb.2023.107809, https://www.sciencedirect.com/science/article/pii/S0747563223001607

55. D.G. Shapiro, R.D. Shachter, User-agent value alignment (2002). https://api.semanticscholar.org/CorpusID:9459542

56. T.B. Sheridan, Human–robot interaction: status and challenges. Hum. Factors **58**(4), 525–532 (2016). https://doi.org/10.1177/0018720816644364. Publisher: SAGE Publications Inc

57. H. Soh, Y. Xie, M. Chen, D. Hsu, Multi-task trust transfer for human–robot interaction. Int. J. Rob. Res. **39**(2–3), 233–249 (2020). https://doi.org/10.1177/0278364919866905

58. N. Wang, D.V. Pynadath, S.G. Hill, N. Wang, D.V. Pynadath, Building Trust in a Human-Robot Team with automatically generated explanations, in *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, vol. 15315 (2015), pp. 1–12

59. N. Wang, D.V. Pynadath, S.G. Hill, Trust calibration within a human-robot team: comparing automatically generated explanations, in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016), pp. 109–116. https://doi.org/10.1109/HRI.2016.7451741

60. A. Washburn, S. Matsumoto, L.D. Riek, Trust-aware control in proximate human-robot teaming, in *Trust in Human-Robot Interaction* (Elsevier, Amsterdam, 2021), pp. 353–377. https://doi.org/10.1016/B978-0-12-819472-0.00015-0, https://linkinghub.elsevier.com/retrieve/pii/B9780128194720000150
61. A. Xu, G. Dudek, Optimo: online probabilistic trust inference model for asymmetric human-robot collaborations, in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2015), pp. 221–228
62. X.J. Yang, V.V. Unhelkar, K. Li, J.A. Shah, 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, 2017), pp. 408–416. https://doi.org/10.1145/2909824.3020230
63. X.J. Yang, Y. Guo, C. Schemanske, 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*, ed. by V.G. Duffy, S.J. Landry, J.D. Lee, N.A. Stanton (2023), pp. 253–265
64. X.J. Yang, C. Schemanske, C. Searle, Toward quantifying trust dynamics: how people adjust their trust after moment-to-moment interaction with automation. Hum. Factors **65**(5), 862–878 (2023). https://doi.org/10.1177/00187208211034716
65. P. Yu, S. Dong, S. Sheng, L. Feng, M. Kwiatkowska, Trust-aware motion planning for human-robot collaboration under distribution temporal logic specifications (2023). http://arxiv.org/abs/2310.01163, arXiv:2310.01163 [cs, eess]
66. L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y.N. Wu, F. Rossano, H. Lu, Y. Zhu, S.C. Zhu, In situ bidirectional human-robot value alignment. Sci. Rob. **7**(68), eabm4183 (2022). https://doi.org/10.1126/scirobotics.abm4183
67. Z. Zahedi, M. Verma, S. Sreedharan, S. Kambhampati, 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, 2023), pp. 281–289. https://doi.org/10.1145/3568162.3578628