Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task

Shreyas Bhat, *Graduate Student Member, IEEE*, Joseph B. Lyons, Cong Shi, and X. Jessie Yang

***Abstract*—In this paper, we present a framework for trust-aware sequential decision-making in a human-robot team wherein the human agent's trust in the robotic agent is dependent on the reward obtained by the team. We model the problem as a finite-horizon Markov Decision Process with the trust of the human on the robot as a state variable. We develop a reward-based performance metric to drive the trust update model, allowing the robotic agent to make trust-aware recommendations. We conduct a human-subject experiment with a total of 45 participants and analyze how the human agent's trust evolves over time. Results show that the proposed trust update model is able to accurately capture the human agent's trust dynamics. Moreover, we cluster the participants' trust dynamics into three categories, namely, Bayesian decision makers, oscillators, and disbelievers, and identify personal characteristics that could be used to predict which type of trust dynamics a person will belong to. We find that the disbelievers are less extroverted, less agreeable, and have lower expectations toward the robotic agent, compared to the Bayesian decision makers and oscillators. The oscillators tend to get significantly more frustrated than the Bayesian decision makers.**

***Index Terms*—Acceptability and trust, human-robot teaming, human-robot collaboration, planning under uncertainty.**

I. Introduction

HUMAN-ROBOT collaborations are becoming more prevalent in a range of fields including package delivery, warehouse management, search and rescue, transportation, and healthcare. To facilitate effective human-robot collaboration, trust has been identified as a key factor. In order to enable trustworthy human-robot interaction, substantial research efforts have been devoted to identifying factors that influence humans' trust in robots [1], developing computational models for trust estimation [2], [3], and developing trust-aware decision making [4], [5].

Manuscript received 23 February 2022; accepted 5 June 2022. Date of publication 6 July 2022; date of current version 18 July 2022. This letter was recommended for publication by Associate Editor A. Clodic and Editor G. Venture upon evaluation of the reviewers' comments. This work was supported by the Air Force Office of Scientific Research under Grant FA9550-20-1-0406. *(Corresponding author: X. Jessie Yang.)*

Shreyas Bhat, Cong Shi, and X. Jessie Yang are with the Industrial and Operations Engineering Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: shreyasb@umich.edu; shicong@umich.edu; xijyang@umich.edu).

Joseph B. Lyons is with the Air Force Research Laboratory, OH 45433-5540 USA (e-mail: joseph.lyons.6@us.af.mil).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at the University of Michigan under Application No. HUM00193551 and HUM00213012, and performed in line with the American Psychological Association code of ethics.

Digital Object Identifier 10.1109/LRA.2022.3188902

Despite these studies, existing research on trust in human-robot interaction is subject to several major limitations. A majority of prior studies focused on trust in automation adopts a *snapshot* view of trust [6]. Trust is measured via questionnaires, typically administered at the end of the experiment. As it is challenging to repeatedly administer trust surveys during the normal functioning of the autonomy, this snapshot view of trust remains inadequate in providing the robotic agent with the moment-to-moment trust the human agent has towards it. More recent developments have tried addressing this research gap via computational models of trust, capable of estimating moment-to-moment changes in trust as a human repeatedly interacts with an autonomous agent [2], [3]. The models in these studies require the definition of a clear binary performance measure, i.e., the autonomy is either correct or wrong. Due to this requirement, these models are used in episodic decision-making scenarios, for example, in a search and rescue scenario where the autonomy detects if a victim is present or not at a site independently of other search sites. In such scenarios, at every site (independent of other sites), the autonomy's performance is considered correct for true positive and true negative detection, and wrong for false positive and false negative detection. However, such models cannot be directly applied in a sequential decision-making scenario wherein the robotic agent needs to perform complex trade off decisions to maximize the cumulative reward and hence the autonomy correctness is more difficult to quantify. In addition, previous studies have revealed the existence of different types of trust dynamics [3], [7]. In [7], the authors found two types of clusters called followers and preservers depending on the trust-dependent behavior of the participants. In [3] the authors found three types of trust dynamics in an *episodic* task setting. We use similar clustering features and find the same clusters in a *sequential* task setting. Moreover, we associate personal characteristics with the type of trust dynamics, which has not been done in prior research.

In this study, we propose an MDP framework with trust of the human on its autonomous partner as a state variable. Incorporating a trust update model allows our autonomous agent to explicitly consider trust and in turn, human behaviour in its decision-making. We introduce a reward-based performance metric to drive the trust estimation algorithm. Finally, from data collected through human-subject experiments, we analyze how human trust evolves with their earned reward over repeated interactions with the autonomy. We find three distinct types of trust dynamics through $k-$means clustering analysis and examine associations between personal characteristics and type of trust dynamics.

2377-3766 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.

The rest of the paper is organized as follows: Section II reviews related work in trust-driven human-robot interaction; Section III formulates the trust-aware decision making problem in an MDP framework and describes the trust-behavior model as well as the trust-seeking reward function; Section IV introduces a reconnaissance mission as a relevant human-autonomy teaming use case to examine how different settings affect the interaction; Section V reports the results and our observations; Section VI summarizes our findings and discusses the limitations and future directions.

II. Related Work

In this section, we discuss three bodies of research motivating the present study.

A. Modeling (Snapshot) Trust and Trust Dynamics

Extensive work has been done in identifying factors that affect (snapshot) trust in automation. These studies often evaluate trust through questionnaires administered at particular instants during the experiment (Refer to [1] for a review of factors). More recently, the research focus has been to construct dynamic models of trust, capable of estimating a human's moment-to-moment trust on an autonomous agent. The primary motivation for such models is their potential to be used to predict real-time human states, thus enabling trust-aware decision making for the autonomous agent. [8] proposed an auto-regressive moving average vector (ARMAV) model to estimate trust at a given time step based on its value at the previous time step, task performance, and whether an automation failure occurred. The Online Probabilistic Trust Inference Model (OPTIMo) proposed in [2] models trust in a performance-centric way as a latent variable in a Dynamic Bayesian Network. In [3], trust is modeled by a Beta distribution with performance induced parameters, personalized for every human.

B. Trust-Aware Planning

The ability for a robot to estimate a human's trust level in real time has led to the development of robots that can adapt their behavior in accordance to trust. In [9], the authors proposed a framework for using estimated trust for Trust-Aware Conservative Control (TACtiC) in which an autonomous agent momentarily changes its behavior whenever the human loses trust. In [10], a trust-workload POMDP is solved to generate optimal policies for a robot to control its transparency to improve the performance of the human-robot team. The authors in [4] propose a trust-POMDP model that can be solved to generate optimal policies for the robot to calibrate the human's trust and improve team performance. [5] presents a reverse psychology model of human trust-behavior and compares it with the more commonly used disuse model.

Most existing literature in trust-aware planning considers trust as a state in an MDP-like problem, wherein the objective of the team is to maximize a reward. Our work has two major departures from previous studies. First, our trust updating rule explicitly depends on the performance of a robot's recommendation, i.e., trust is increased when the actual performance (reward) of the robot's recommendation exceeds that of the opposite recommendation. This performance-based trust updating rule is new and more practical. Second, our immediate reward has a trust gaining term incentivizing the robot to make righteous recommendations, so that trust is positively reinforced.

C. Individual Differences in (Snapshot) Trust and Trust Dynamics

Most existing research investigating individual differences aims to find associations between individual characteristics and (snapshot) trust. Individuals with high propensity to trust automation had a higher difference in post-task trust between a reliable and faulty automation [11]. In addition, the human agent's personality trait of neuroticism was found to be negatively correlated with agreement with automation [12]. The Perfect Automation Schema, which represents one's dispositional expectations of technology [13], has been shown to be a significant factor affecting post-task trust in automation [14]. It is possible that these individual differences will be associated with different types to trust dynamics.

III. Problem Formulation

In this section, we describe the framework for modeling and incorporating trust in the decision-making system of a robotic agent. The agent provides recommendations to their human partner about the action that they should take, but the final decision of action selection lies with the human. We use the trust dynamics model described in [3] with some modifications to suit our problem. The recommendation system of the agent is modeled as a finite horizon Markov Decision Process (MDP) with trust as a state variable. The agent solves an optimization problem to maximize the expected cumulative future reward. The specific scenario we target is an 'Intelligence, Surveillance, and Reconnaissance' (ISR) mission in which a human soldier teams up with an *intelligent* drone to search through a town for the presence of threats. The recommendation system guides the soldier on whether s/he should breach a site directly or deploy a Robotic Armored Rescue Vehicle (RARV). Using the RARV prevents any health loss to the soldier in the presence of a threat, but it takes additional time to deploy the RARV each time. On the other hand, breaching a site directly is faster, but the soldier will be harmed if a threat is present inside the site. Here, two natural (but conflicting) goals that arise are to minimize any damage to the soldier while also minimizing the time to search through all the sites.

A. Trust Aware Decision-Making

We model the ISR task as a trust-aware MDP (Fig. 1) with the objective of the team being to minimize a weighted sum of the health loss to the soldier and the time taken to complete the mission. A trust-aware MDP is a tuple of the form (S, A, H, T, R), where S is a set of states, A is a set of actions for the autonomous agent, H is the embedded human behavior model, $T(s, a)$ is the transition function, and $R(s, a)$ is a reward function that the agent tries to maximize. Details of the definition of our MDP are given below.

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.

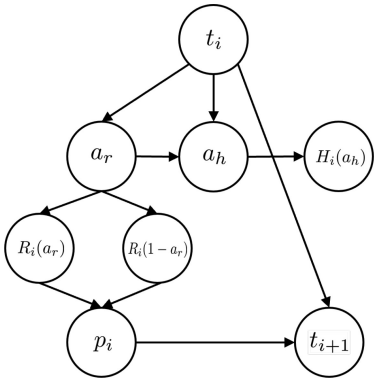


Fig. 1. Graphical Representation of our MDP.

1) States: We use the estimated trust of the human on the robot as the state variable. More specifically, a state is specified by a tuple (α, β). Details of how this translates to the trust of the human are given in the description of transition function below.

2) Actions: At each search site, the two actions available to the autonomous agent are to either recommend to use the RARV or to recommend to not use the RARV.

3) Human Behavior Model: The embedded human behavior model encodes how a human agent responds to actions (or recommendations) made by an autonomous agent. We assume 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 selects the opposite action of the one that was recommended. More precisely, let a_r and a_h denote the action taken by the autonomous agent and the human agent, respectively.

$$\begin{aligned} \mathbb{P}(a_h = a_r) &= t_i, \\ \mathbb{P}(a_h = 1 - a_r) &= 1 - t_i. \end{aligned} \tag{1}$$

Here t_i is the trust level at the i^{th} search site.

4) Reward Function: The reward function uses a weighted average of health loss and time. Let h and c be the (constant) health and time losses, respectively. Let w_h and w_c be the weights for the health and time losses, respectively. We define $H_i(a_h)$ to be the immediate reward given that the human chooses action a_h at site i. Then

$$E[H_i(a_h)] = -(1 - a_h)\hat{d}_i w_h h - a_h w_c c, \tag{2}$$

where $\hat{d}_i$ is the probability of threat presence at site i.

Then, we define $R_i(a_r)$ to be the immediate *task* reward at site i, given that the robot's recommendation is a_r. Now, by *fixing* the robot's recommendation to $a_r = a$, via our human behavior model defined in (1), we have

$$\begin{aligned} & E[R_i(a_r = a)] \\ &= E[E[R_i(a_r = a)|t_i]] \\ &= E[t_i E[H_i(a_h = a)] + (1 - t_i)E[H_i(a_h = 1 - a)]] \\ &= \hat{t}_i E[H_i(a_h = a)] + (1 - \hat{t}_i)E[H_i(a_h = 1 - a)], \end{aligned} \tag{3}$$

where the second equality is due to (1) and the third equality holds because $E[H_i(a_h = a)]$ is a constant given a fixed a.

It was noted in simulations that if we only use this *task* based reward, the autonomous agent learns that for most people, trust decreases after failures more easily than increasing after successes [3], and thus it exploits this human behavior by always recommending the opposite action. Trust being low, the human chooses the action opposite to the recommendation, thus increasing performance. We call this behavior of the autonomy as *reverse-psychology*. Since such deceptive behavior is undesirable, we add a trust gain reward to the reward function to incentivize the autonomous agent to make righteous recommendations. Therefore, we define the expected immediate reward at site i with robot's recommendation a as

$$E[\mathbf{IR}_i^a] = E[R_i(a_r = a)] + E[\underbrace{\lambda_i \cdot \mathbb{1}(A)}_{G_i(a_r = a)}], \tag{4}$$

where the weight for "trust-gain" is given by

$$\lambda_i = w_t \sqrt{N - i}. \tag{5}$$

The term $G_i(a_r = a)$ represents the trust gain reward at site i. We define A as the event when trust increases, i.e., $\mathbb{1}(A) = 1$ if the performance of the autonomous agent is a success and $\mathbb{1}(A) = 0$ otherwise (Note: we define our notion of performance via (10)). The parameter λ_i is a weight given to the trust gain reward that decreases with the stage number. The idea behind this is to support trust-gaining behavior near the current stage, and performance optimizing behavior towards the later stages of planning.

In our experiments, we used the following values for the constants, $h = 100, c = 150, w_h = 0.85, w_c = 0.15, w_t = 10$. We chose these values as they resulted in about 80% threat detection accuracy for the drone in simulations.

5) Transition Function: We modify the model described in [3] to use a binary *reward-based performance* for the trust update step. We fit personalized trust parameters $(\alpha_0, \beta_0, w^s, w^f)$ for each participant to model their trust.

$$t_i \sim Beta(\alpha_i, \beta_i), \tag{6}$$

$$\hat{t}_i = E[t_i] = \frac{\alpha_i}{\alpha_i + \beta_i}. \tag{7}$$

Here, t_i is the trust level at the i^{th} stage. The parameters α_i and β_i are updated as follows.

$$\alpha_i = \begin{cases} \alpha_{i-1} + w^s, & \text{if } P_i = 1, \\ \alpha_{i-1}, & \text{if } P_i = 0. \end{cases} \tag{8}$$

$$\beta_i = \begin{cases} \beta_{i-1}, & \text{if } P_i = 1, \\ \beta_{i-1} + w^f, & \text{if } P_i = 0. \end{cases} \tag{9}$$

The intelligent agent has two (conflicting) goals: to minimize any damage to the soldier and to minimize the time to search through all the sites. Since the autonomous agent's optimal recommendation comes from a reward maximization factoring in both goals, its performance cannot directly be judged by whether its recommendation matches with the presence of threat

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.

inside a building. To overcome this, we define the performance of the drone based on immediate rewards earned by the participant.

$$P_i = \begin{cases} 1 & \text{if } R_i(a_r) \geq R_i(1-a_r), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

At each site i, we observe the presence of threat and compute the realizations of the reward for following the recommendation $R_i(a_r)$, and that for doing the opposite $R_i(1-a_r)$. We compute the performance of the drone based on these values.

In our experiment, the participants report their trust level on the robot after each interaction (see section IV for details). We use gradient descent on the Bayesian posterior given the performance history and the prior (from [3]) over the model parameters to update our estimates of $(\alpha_0, \beta_0, w^s, w^f)$ in real time after receiving this feedback from the participant. We use the digamma function approximation presented in [15] to approximate the gradients of the beta distribution function.

6) Value Iteration: We solve the MDP using value iteration on the Bellman Equations. The action at the current state is selected by maximizing the expected reward at the current step summed together with a discounted value of the next state at the next stage.

$$V_i^a = E\left[\mathbf{IR}_i^a\right] + \sum_{(\alpha_{i+1},\beta_{i+1})\in S} \gamma \mathbb{P}(\alpha_{i+1}, \beta_{i+1} | \alpha_i, \beta_i, a) V_{i+1}. \tag{11}$$

$$V_i = \max_a V_i^a. \tag{12}$$

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

$$V_N = \max_a E\left[\mathbf{IR}_i^a\right]. \tag{13}$$

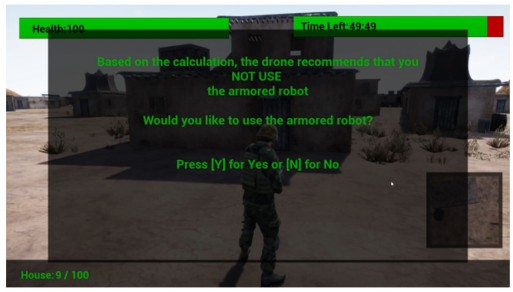
IV. Experiment

This section describes details of the human-subject experiment. The experiment complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board at the University of Michigan.

A. Participants

A total of 46 adults participated in the study. One participant's data was discarded as the participant marked all survey questions in the middle and used significantly less time compared to other participants. The remaining 45 participants consisted of 21 females and 24 males (Age: Mean = 22.8 years, *SD* = 3.6 years). The participants were recruited over two phases. In the first phase, we had 31 participants. We modelled their trust dynamics using the model described in Section III. Clustering of the trust dynamics revealed three types with 20, 3, 8 participants respectively. As the numbers of participants in types 2 and 3 were small for statistical analysis, we recruited another 14 participants. For the second phase of participant recruitment, all potential participants filled a pre-experimental survey (see Section IV-C for details) and we selected the ones whose personal characteristics profile were similar to participants identified as type 2 and type 3 in the first phase. Each participant was reimbursed with a base pay of $20 with a bonus of up to $10 based on their performance on the task. The performance was measured by the time taken by the participants to complete the task and the final health level of the soldier.

(a) A view of the 3d environment of the testbed



(b) An example of the recommendation GUI

Fig. 2. The testbed developed in the Unreal Engine 4 game engine.

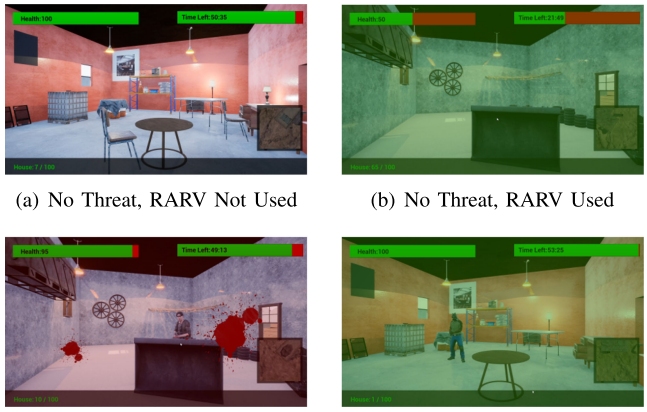
B. Testbed

We developed a 3D testbed using the Unreal Engine game development platform. A screenshot of the testbed is shown in Fig. 2(a). It shows the *intelligent* drone, the soldier, and the RARV. Fig. 2(b) shows a screenshot of the recommendation dialog box where the participant was recommended to not use the RARV. It also shows the two bars showing the health level of the soldier and the time remaining to complete the mission, emphasizing the two objectives the participants need to optimize. Once the participant makes a choice of action, the four possibilities depending on the presence of threat and participant's selected action are shown in Fig. 3. The participants are told that each time they encounter a threat without using the RARV, they will lose 5 points of health while deploying the RARV takes approximately 10 seconds. They are told to choose an action based on their interaction history and the recommendation from the drone. After exiting each house, the participants are asked to adjust a slider to give feedback on their level of trust on the drone's recommendations.

C. Measures

1) Personality: The big 5 factors of personality (Extraversion, Conscientiousness, Agreeableness, Neuroticism, and Imagination) were measured using the 20-item mini-IPIP scale [16]. This 5-point Likert scale has widely been used in human-robot trust research [17]. This survey was administered pre-experiment.

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.



(a) No Threat, RARV Not Used (b) No Threat, RARV Used

(c) Threat, RARV Not Used (d) Threat, RARV Used

Fig. 3. The four outcomes based on the presence of threat inside a site and the choice of action by the participant.

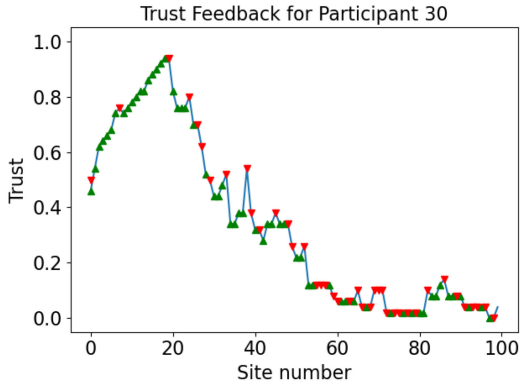


Fig. 4. Trust feedback with overlay of red and green triangles representing the rewards. Green triangles show that the participant would receive a higher reward by following the recommendation and red triangles show that the participant would receive a higher reward by not following the recommendation.

2) Perfect Automation Schema: Perfect Automation Schema (PAS) was measured using the 7-item scale developed by [13]. Of these, 4 items measure high expectations from the autonomy and the other 3 items measure All-or-none thinking. This was a 7-point Likert scale. This survey was administered pre-experiment.

3) Propensity to Trust: Propensity to trust autonomous systems was assessed using a 6-item scale developed in [18]. This was also a 5-point Likert scale. This survey was administered pre-experiment.

4) Moment-to-Moment Trust: During the task, the participants were asked to rate their moment-to-moment trust on the drone by adjusting a slider on a 100-point scale.

5) Post-Experiment Trust: After the experiment, we used two scales [17], [19] for assessing trust. The first one was a 8-item questionnaire with sliders while we used 6-items from the second scale all of which were 7-point Likert type questions.

6) Workload: Workload was measured using the NASA Task Load Index [20] administered after the experiment. We only used 5 of the 6 items as there was no physical demand from the participants in our experiment. All the items had the participants rate their feelings on a slider with values ranging from Very low to Very high.

D. Experimental Procedure

Prior to the experiment, participants provided informed consent and completed several surveys assessing their demographics, personality traits, perfect automation schema, and propensity to trust automation. They were oriented to the steps of the experiment and walked through each of the screens they would see during the experiment. The two-fold objective of minimizing time and maximizing health was emphasized. Participants were told that the robotic agent was imperfect, but they were not informed of the exact reliability level. They were also told that the robotic agent's recommendations would help them achieve the two-fold objective (i.e., to minimize any damage to the soldier and to minimize the time to search through all the sites). They were informed about the performance-based bonus pay. Participants then proceeded to the experimental trials, wherein they had to search through 100 houses sequentially. After searching each house, the participants were asked to report their level of trust on the autonomous agent's recommendations. The participants took on average 51.5 minutes to complete the task. The average threat detection accuracy of the robotic agent was around 85% across the participants. At the end of the experiment, participants reported their post-experiment trust toward the automated aid using two scales. They also reported their level of workload.

V. Results & Discussion

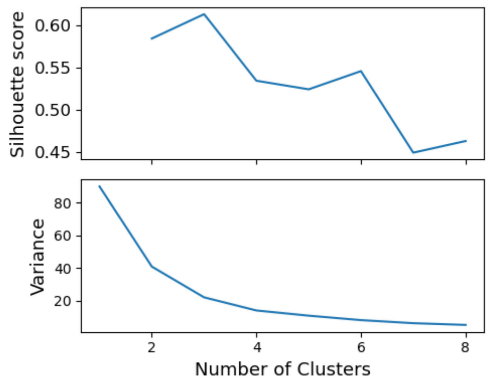
A. Using Immediate Actual Reward as a Performance Metric

Since the participants are explicitly told to consider both the soldier's health and the time to complete the search as their objectives, we expect their trust to be correlated with the immediate reward that they receive upon choosing an action. We expect that a participant's trust would be likely to increase if following the recommendation by the drone gets them a higher reward that not following the recommendation and vice versa. Fig. 4 shows a representative example. In the figure, a green triangle represents the site at which following the recommendation would result in a better immediate task reward gain $(P_i = 1)$ and the red triangles represent the opposite. It is quite clear that a red triangle is often followed by a decrease in trust and a green triangle is followed by an increase in trust. Thus, this reward-based performance metric is able to capture moment-to-moment trust changes of the participant. Using this performance metric in our trust update model, we get a prediction root mean squared error of 0.1266 $(SD = 0.078)$ across the participants (Note: The reported trust values were between 0 and 1).

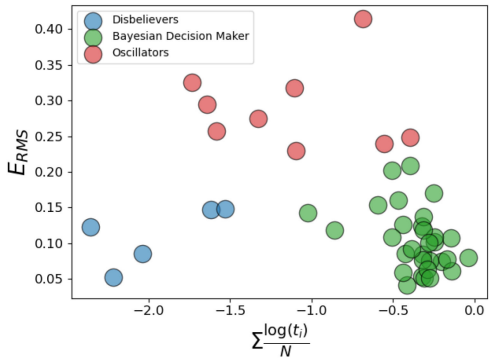
B. Clustering of Trust Dynamics

We employ k-means clustering to group together participants with similar trust dynamics. Bench marking a prior study [3], we use two features, namely the average logarithm of trust, and the E_{RMS}, for the clustering analysis. The average logarithm of trust represents the overall trust a human agent has on the robotic agent, which is computed as the logarithm of the trust feedback

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.



(a) Variances and Silhouette Scores



(b) Clusters

Fig. 5. Clustering of participants according to their trust dynamics. The features used are the root mean squared error between our predictions and their feedback and their average log trust. As evidenced by the elbow in the variance plot and the maximum in the silhouette scores plot, we chose $k = 3$ as the optimum number of clusters. There are 31 Bayesian Decision Makers, 5 Disbelievers and 9 Oscillators.

values over the 100 trials. The RMS error (E_{RMS}) represents the extent to which a human agent's trust updating process is Bayesian. For computing E_{RMS}, we consider the feedback at the first 20 sites as a training set. Thereafter, we use the feedback after every 5 sites to update the parameters of our model. Thus, we compute the E_{RMS} over the last 80 sites. Fig. 5 shows the results of the k-means algorithm. We note elbows in the variance plot at $k = 2$ and $k = 3$ clusters. Employing the silhouette score metric, we see that there is a peak at $k = 3$. We thus choose $k = 3$ as the optimum number of clusters. One of the clusters includes participants with small E_{RMS} and generally higher values of trust. We call this cluster *Bayesian Decision Makers* as their trust is well predicted by our model. The second significant cluster consists of participants whose E_{RMS} values are small but whose trust is generally low. We call this cluster *Disbelievers* because of their low trust value irrespective of the performance of the agent. The third significant group is one with high E_{RMS} values, whose trust cannot be predicted very well with our model. Their trust changes very rapidly from moment-to-moment, thus making it harder to predict. Representative plots of each of these clusters can be seen in Fig. 6. Even though we lack the "true" cluster labels, prior empirical studies examining trust in automation and human decision making provide face and external validity

TABLE I
MEAN AND STANDARD DEVIATION (SD) OF PERSONAL CHARACTERISTICS BETWEEN THE THREE DIFFERENT TRUST DYNAMICS (BDM = BAYESIAN DECISION MAKER)

Personal Characteristic	BDM	Disbeliever	Oscillator
Extraversion (/20) *	9.5 (3.3)	5.8 (2.8)	11.3 (2.9)
Agreeableness (/20) *	13.5 (2.5)	10.4 (5.0)	14.1 (1.8)
Conscientiousness (/20)	13.1 (2.7)	12.4 (3.0)	12.1 (4.5)
Neuroticism (/20)	7.9 (2.7)	6.8 (3.6)	10.2 (4.7)
Intellect/Imagination (/20) †	11.7 (2.0)	9.8 (1.8)	12.2 (1.8)
High Expectations (/28) **	12.7 (3.9)	6.4 (2.8)	12.4 (4.2)
All or None Thinking (/21)	6.6 (2.9)	6.4 (3.4)	7.1 (3.1)
Trust Propensity (/30) †	20.2 (4.4)	17.2 (4.1)	22.8 (3.2)

$**-p < 0.01$, $*-p < 0.05$, $\dagger-p < 0.1$

TABLE II
MEAN AND STANDARD DEVIATION (SD) OF POST EXPERIMENT METRICS BETWEEN THE THREE DIFFERENT TRUST DYNAMICS

Personal Characteristic	BDM	Disbeliever	Oscillator
Trust (Muir) (/100) ***	65.4 (13.5)	15.8 (9.9)	44.7 (26.1)
Trust (Lyons) (/7) ***	4.5 (0.54)	3.1 (0.6)	3.6 (0.9)
Mental Demand (/100)	39.6 (25.2)	42.0 (36.6)	50.3 (28.6)
Temporal Demand (/100)	50.8 (27.4)	62.0 (24.5)	42.9 (21.3)
Performance (/100)	58.6 (19.7)	50.8 (30.0)	46.2 (31.7)
Effort (/100)	34.3 (23.0)	34.4 (17.4)	49.8 (32.2)
Frustration (/100) *	45.8 (22.2)	58.4 (25.4)	68.1 (14.3)

$***-p < 0.001$, $*-p < 0.05$

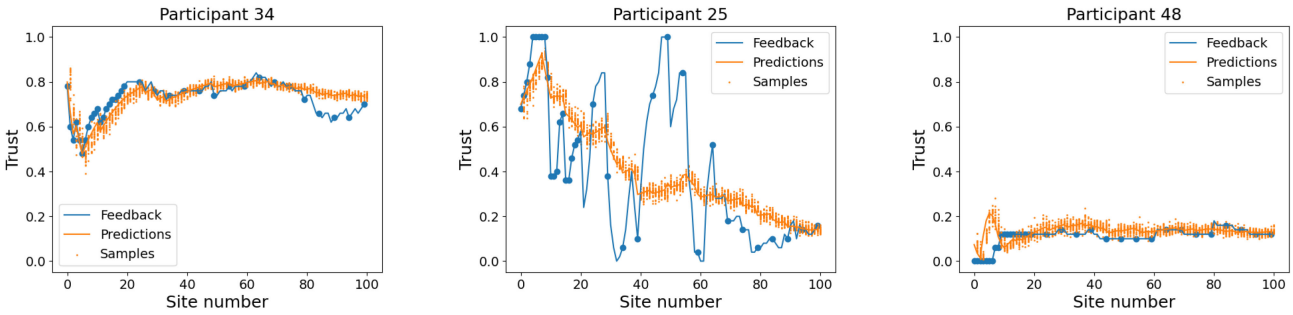
for the three types of trust dynamics. Disbelievers have been reported in studies investigating public's trust and acceptance of automated driving - around 40% of a JD power survey correspondents said they "would not ride in an (automated vehicle) *regardless of* what progress is made [21]." Bayesian thinkers, who form accurate prior probabilities and update their belief based on new information, have been observed in everyday cognitive judgements [22]. In addition, the oscillators could be considered as Bayesian thinkers with bounded rationality, who only update their belief based on immediate past history, resulting in significant fluctuation in trust assessment.

We find smaller E_{RMS} for the Bayesian Decision Makers (Mean = 0.093 and SD = 0.04) and Disbelievers (Mean = 0.1 and SD = 0.04), compared to the Oscillators (Mean = 0.26 and SD = 0.05), suggesting our trust dynamics model is able to accurately represent the dynamics of Bayesian Decision Makers and Disbelievers, but is unable to represent the dynamics of Oscillators very well.

C. Association Between Personal Characteristics and Type of Trust Dynamics

In this section, we present significant individual differences between the three identified clusters of trust dynamics. Our findings are summarized in Table I and II. $p < .05$ is considered significance. We performed one-way ANOVA between the data of the clustered participants. Results showed significant difference between the three types of trust dynamics in Extraversion ($F(2, 42) = 4.991, p = 0.011$),

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.



(a) Bayesian Decision Maker (b) Oscillator (c) Disbeliever

Fig. 6. Three types of trust dynamics over an interaction period of 100 sites. The blue curve shows the reported trust feedback and the orange curve shows the model predicted trust. The blue points represent the sites at which reported feedback was used to train the model. Finally, the orange dots represent points sampled from the model's beta distribution.

Agreeableness $(F(2, 42) = 3.276, p = 0.048)$, and the High Expectations facet of the Perfect Automation Schema $(F(2, 42) = 5.752, p = 0.006)$. Further, propensity to trust automation $(F(2, 42) = 3.002, p = 0.06)$ and intellect/imagination $(F(2, 42) = 2.687, p = 0.08)$ seemed to be different. However, these differences did not reach significance.

Post-hoc analysis with Bonferroni adjustment test shows that there is a significant difference in Extraversion between the Oscillators and the Disbelievers $(p = 0.009)$, with the disbelievers being significantly less extroverted than the oscillators. There seemed to be a trend that the Bayesian Decision Makers were more extroverted than the Disbelievers $(p = 0.061)$. However the trend did not reach significance. Disbelievers had significantly lower expectations from automation compared to both bayesian decision makers $(p = 0.005)$ and oscillators $(p = 0.023)$. In the case of Agreeableness, there seemed to be trends that the disbelievers were less agreeable than the other two groups ($p = 0.068$ between disbelievers and bayesian decision makers and $p = 0.06$ between disbelievers and oscillators). The trends, unfortunately, did not reach significance.

Performing one way ANOVA on the post-experiment measures show significant difference between the three types of trust dynamics in their post-experiment trust reports (Trust questionnaire by Muir and Moray $F(2, 42) = 22.167, p < 0.001$, Trust questionnaire by Lyons and Guznov $F(2, 42) = 15.183, p < 0.001$) and their frustration levels $(F(2, 42) = 4.136, p = 0.023)$).

Post-hoc analysis with Bonferroni adjustment shows that there are significant differences between each of the three groups' trust reports according to the trust questionnaire by Muir and Moray ($p < 0.001$ between bayesian decision makers and disbelievers, $p = 0.006$ between bayesian decision makers and oscillators, and $p = 0.009$ between osciallators and disbelievers with the highest trust for bayesian decision makers and lowest for disbelievers). Trust reported with the trust questionnaire by Lyons and Guznov only showed significant difference between disbelievers and bayesian decision makers $(p < 0.001)$, and between oscillators and bayesian decision makers $(p = 0.002)$. We believe that a couple of reasons might be behind this difference. One, the oscillators may be inherently unsure about their level of trust on the robot, causing the high variance in their moment-to-moment trust reports. Ultimately, the oscillators may have experienced a state of suspicion in relation to the drone aid [23]. Suspicion represents the combination of high cognitive activity, high uncertainty, and perceived malicious intent [24]. A perusal of the workload data does appear to suggest that the oscillators experienced higher workload relative to the other clusters. This may have been indicative of suspicion for the oscillators whereas the Bayesian decision makers and the disbelievers were more certain of their trust and distrust, respectively. Secondly, the trust measures had some conceptual differences. The Muir and Moray measure focused on elements of trustworthiness (competence, reliability, faith, dependability, etc.) while the Lyons and Guznov measure focused on one's willingness to be vulnerable to the drone aid. It is clear that the three clusters were sensitive to variations in trustworthiness, yet it appears that only Bayesian decision makers were willing to be vulnerable to the drone aid. This has interesting implications for trust modeling given that one's intentions to be vulnerable are a precursor to risk taking.

Results also showed that oscillators were significantly more frustrated than bayesian decision makers $(p = 0.025)$

VI. Conclusion

In this study, we developed a framework to explicitly incorporate trust in the decision-making system of an autonomous recommendation system. We formulated the problem of finding the optimal recommendation as an MDP with an objective function with performance-maximizing and trust-gaining components. We demonstrated the presence of three distinct types of trust dynamics - 1) Bayesian Decision Makers, who vary their trust according to the performance of the autonomy and whose trust stabilizes after repeated interactions, 2) Oscillators, whose trust varies wildly after each interaction, with little to no stabilization even after many interactions, and 3) Disbelievers, whose trust in autonomy is low irrespective of the autonomy's performance. Our trust estimation model is able to predict the trust states of Bayesian Decision Makers and Disbelievers with good accuracy. However, it is not able to capture the moment-to-moment variation of trust of an Oscillator accurately, thus pointing to a requirement to use a different model for people belonging to this category.

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.

Given the value of establishing a robust and dynamic model of trust among individuals interacting with a machine partner, the current study suggests that there may be merit in using individual differences and state measures as a basis to evaluate the feasibility of using such methods. In particular, it would be advantageous to identify who would fall into the three taxonomies of trust dynamics. The results suggest that those individuals classified as Bayesian Decision Makers evidenced high expectations of automation. When combined with other state measures such as trust and frustration, one might be able to identify a profile of the Bayesian Decision Makers as the combination of individual differences and state measures could be used to parse the sample into the three categories of trust dynamics. Knowing that an individual might fall into one of the categories could influence whether or not a machine partner that is equipped with a dynamic trust model is a feasible solution for that individual.

Results of the study should be viewed in light of the following limitations. First, we assume that the human behaves according to a reverse psychology model. Incorporating a different human trust-behavior model in our MDP formulation could be a direction for future work. Second, we currently assume that the weights in the reward function are fixed. However, these weights could vary from person to person. For example, one person can be very concerned with protecting the soldier's health, thus having a large w_h and another person could be more concerned with taking as little time as possible to complete the search, thus having a larger w_c. It can also be possible that the weights are a function of the current health and time. Future research can use our framework, in conjunction with inverse reinforcement learning [25], [26] to define personalized weights in the reward function, thus having personalized metrics for performance of the autonomous agent. Third, we used two features in the clustering analysis. Although these two features were based on prior literature, additional features, such as human agents' decision making biases [6], could be considered to build high-dimension clusters in future studies. Fourth, as the number of participants in each cluster was determined post-hoc instead of *a priori*, the sample size was unbalanced. As far as we know, this is the first study to explore the association between personal characteristics and trust dynamics. Utilizing the identified associations, future research could create a more balanced dataset with higher statistical power.

REFERENCES

[1] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, "Evolving trust in robots: Specification through sequential and comparative meta-analyses," *Hum. Factors*, vol. 63, no. 7, pp. 1196–1229, 2021.

[2] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proc. IEEE/ACM 10th Int. Conf. Hum.-Robot Interaction (HRI)*, 2015, pp. 221–228.

[3] Y. Guo and X. J. Yang, "Modeling and predicting trust dynamics in human-robot teaming: A bayesian inference approach," *Int. J. Social Robot.*, vol. 13, pp. 1899–1909, 2021.

[4] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-aware decision making for human-robot collaboration: Model learning and planning," *J. Hum.-Robot Interact.*, vol. 9, no. 2, pp. 1–3, Jan. 2020. [Online]. Available: https://doi.org/10.1145/3359616

[5] Y. Guo, C. Shi, and X. J. Yang, "Reverse psychology in trust-aware human-robot interaction," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 4851–4858, Jul. 2021.

[6] X. J. Yang, C. Schemanske, and C. Searle, "Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation," *Hum. Factors*, 2021, Art. no. 00187208211034716.

[7] G. McMahon, K. Akash, T. Reid, and N. Jain, "On modeling human trust in automation: Identifying distinct dynamics through clustering of markovian models," *IFAC-PapersOnLine*, vol. 53, pp. 356–363, 2020.

[8] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[9] A. Xu and G. Dudek, "Maintaining efficient collaboration with trust-seeking robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2016, pp. 3312–3319.

[10] K. Akash, T. Reid, and N. Jain, "Improving human-machine collaboration through transparency-based feedback - part II: Control design and synthesis," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 322–328, 2019, 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS, 2018.

[11] S. Merritt and D. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Hum. factors*, vol. 50, pp. 194–210, 05 2008.

[12] J. L. Szalma and G. S. Taylor, "Individual differences in response to automation: The five factor model of personality," *J. Exp. Psychol. Appl.*, vol. 172, pp. 71–96, 2011.

[13] S. M. Merritt, J. L. Unnerstall, D. Lee, and K. Huber, "Measuring individual differences in the perfect automation schema," *Hum. Factors*, vol. 57, no. 5, pp. 740–753, 2015.

[14] J. Lyons and S. Guznov, "Individual differences in human-machine trust: A multi-study look at the perfect automation schema," *Theor. Issues Ergonom. Sci.*, vol. 20, pp. 440–458,2019.

[15] C. Mortici, "The proof of muqattash-yahdi conjecture," *Math. Comput. Modelling*, vol. 51, pp. 1154–1159,2010.

[16] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality," *Psychol. Assessment*, vol. 18, no. 2, pp. 192–203, 2006.

[17] J. B. Lyons, C. S. Nam, S. A. Jessup, T. Q. Vo, and K. T. Wynne, "The role of individual differences as predictors of trust in autonomous security robots," in *Proc. IEEE Int. Conf. Hum.- Mach. Syst. (ICHMS)*, 2020, pp. 1–5.

[18] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Hum. Factors*, vol. 55, no. 3, pp. 520–534, 2013.

[19] B. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, pp. 429–460, 1996.

[20] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, vol. 52, pp. 139–183.

[21] J. D. Power, "Automated vehicles and insurancc pulse policy," *J. D. Power, Tech. Rep.*, 2018. [Online]. Available: https://www.namic.org/pdf/18memberadvisory/181008_Automated_Vehicles_JD_Power_NAMIC_Questionnaire.pdf

[22] T. L. Griffiths and J. B. Tenenbaum, "Optimal predictions in everyday cognition," *Psychol. Sci.*, vol. 17, no. 9, pp. 767–773, 2006.

[23] J. B. Lyons, C. K. Stokes, K. J. Eschleman, G. M. Alarcon, and A. J. Barelka, "Trustworthiness and IT suspicion: An evaluation of the nomological network," *Hum. Factors*, vol. 53, no. 3, pp. 219–229, 2011.

[24] P. Bobko, A. J. Barelka, and L. M. Hirshfield, "The construct of state-level suspicion: A. model and research agenda for automated and information technology (IT) contexts," *Hum. Factors*, vol. 56, no. 3, pp. 489–508, 2014.

[25] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proc. 17th Int. Conf. Mach. Learn.*, Morgan Kaufmann, 2000, pp. 663–670.

[26] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 3916–3924.

Authorized licensed use limited to: University of Michigan Library. Downloaded on April 13,2025 at 20:03:15 UTC from IEEE Xplore. Restrictions apply.