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

ICRA 2022
IEEE International Conference on Robotics and Automation

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

UNIVERSITY OF MICHIGAN

Motivation

- Previous studies [1], [2] have revealed the existence of distinct types of trust dynamics, but have not tried to associate personal characteristics with the type of trust dynamics
- Most computational models of trust [2], [3] require the definition of a binary performance metric for the autonomy
- Such a performance metric is difficult to define in a sequential decision-making task where the goal of the autonomy is to maximize the cumulative reward.

Problem Formulation

We propose a finite horizon Markov Decision Process (MDP) for modeling and incorporating trust in the decision-making system of a robotic agent

A trust-aware MDP is a tuple of the form (S, A, H, T, R)

- S is a set of states
- A is a set of actions
- H is the embedded human behavior model
- $T(s, a)$ is the transition function
- $R(s, a)$ is the reward function

We target the specific scenario in which the human-autonomy team sequentially search through houses in a town for threats.

Here, a state is represented by parameters (α, β) which represents the trust level,

$$t \sim Beta(\alpha, \beta)$$

the actions for the autonomy are whether to recommend to use protective measures or whether to breach a house directly, the human behavior is modeled via the trust level, namely,

$$P(a_h = a_r) = t_i$$
$$P(a_h = 1 - a_r) = 1 - t_i$$

the (negative) rewards are a weighted sum of the health loss cost and the time loss cost,

$$R_i(a) = -w_h h(a) - w_c c(a)$$

and the transition function represents the trust update model,

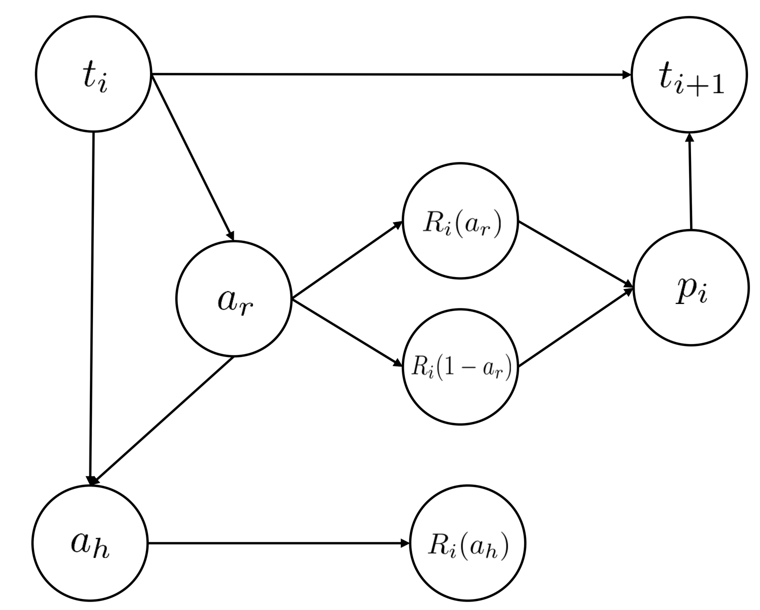
$$\alpha_i = \begin{cases} \alpha_{i-1} + w^s, & \text{if } p_i = 1, \\ \alpha_{i-1}, & \text{if } p_i = 0. \end{cases}$$

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

with the *reward-based* performance metric,

$$p_i = \begin{cases} 1 & \text{if } \; R_i(a_r) \geq R_i(1 - a_r) \\ 0 & \text{otherwise} \end{cases}$$

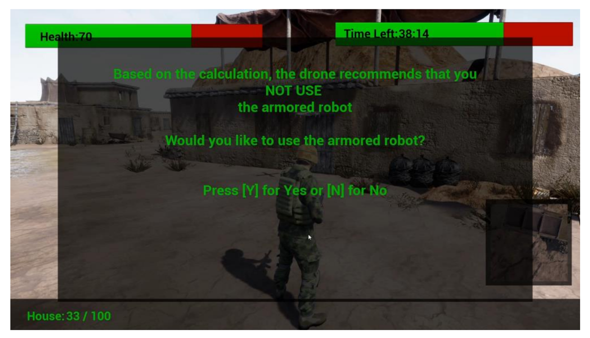
The trust-aware MDP represented graphically



Testbed

We developed a 3D testbed in the Unreal Engine. The intelligent drone recommends whether to use or not use a Robotic Armored Rescue Vehicle (RARV). The participant has the final choice on whether to use the RARV.

We conducted a human-subject study with 45 participants. The results are presented below.



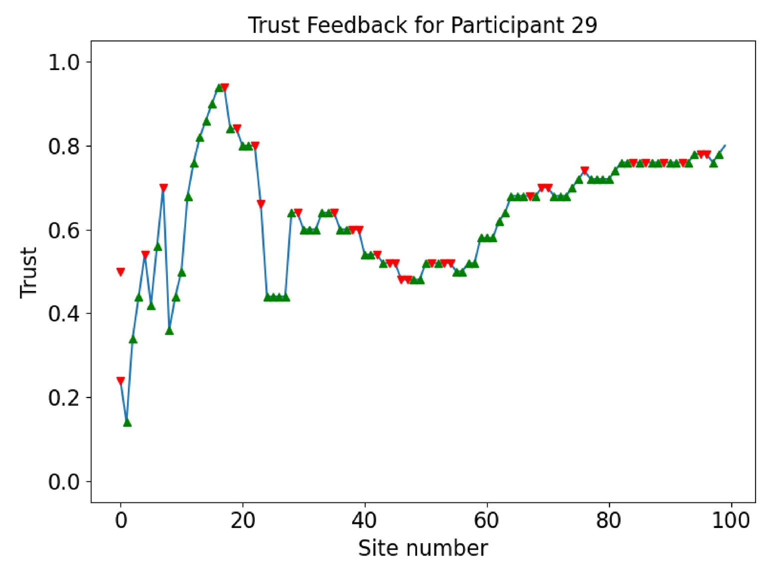
Example of the GUI

Example scenario on choosing an action

Immediate Reward as a Performance Metric

We see that trust is correlated with the immediate reward that they receive upon choosing an action.

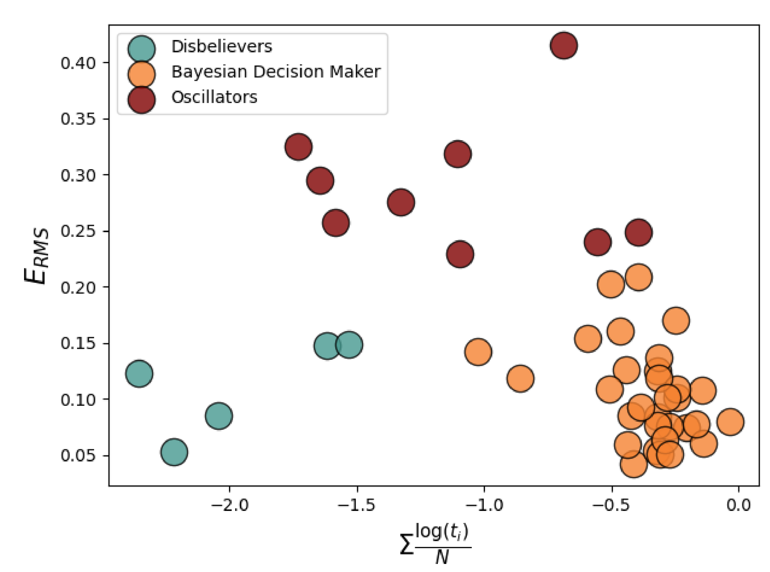
A positive reward-performance (green triangle) usually leads to an increase in trust and vice versa



Clustering Trust Dynamics

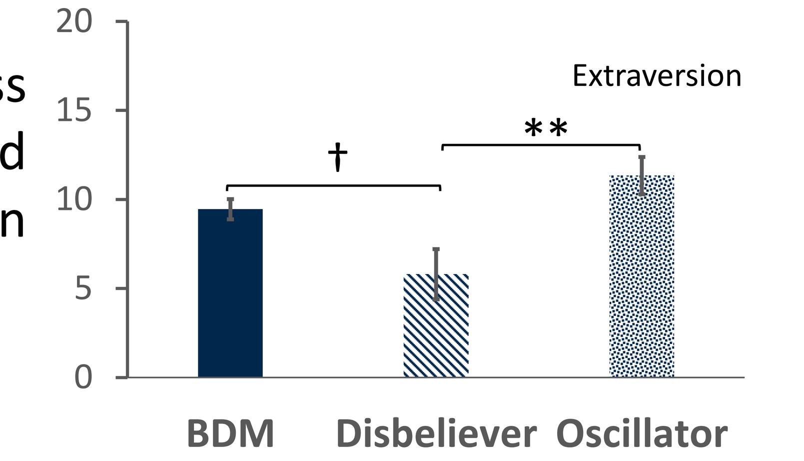
We used k-means clustering to group participants with similar trust dynamics with the root mean squared error and the average log trust as features.

We found three significant clusters – Bayesian Decision Makers, Disbelievers, and Oscillators

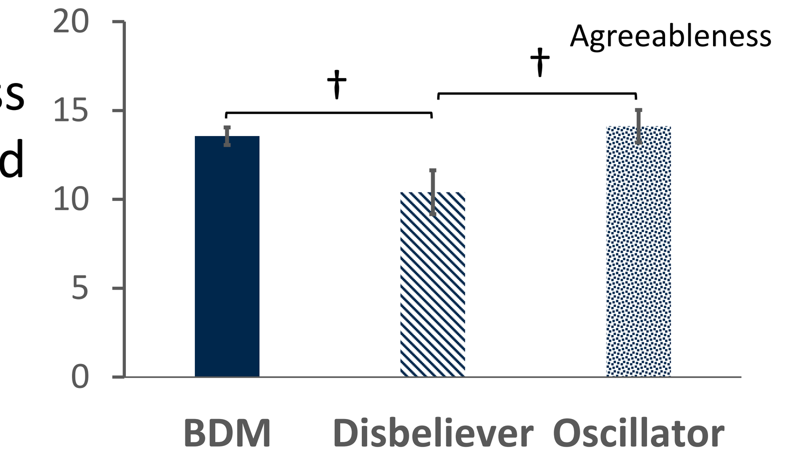


Personal Traits and Types of Trust Dynamics

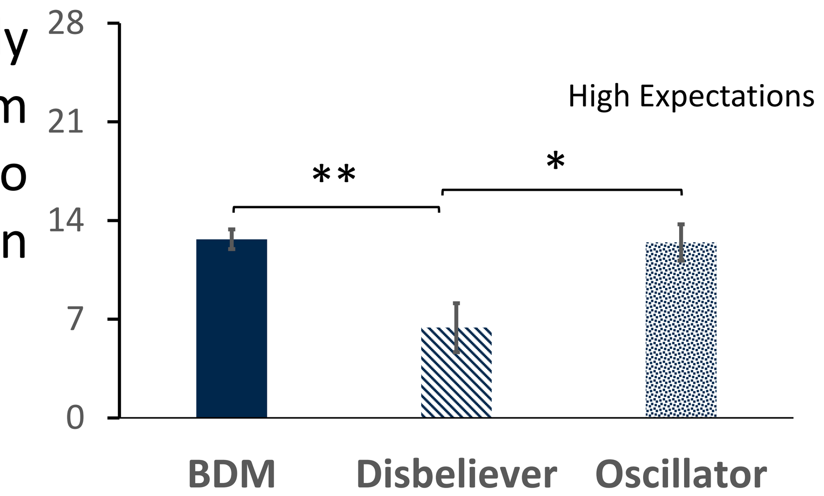
Disbelievers significantly less extroverted than Oscillators and marginally less extroverted than Bayesian Decision Makers



Disbelievers marginally less agreeable than Oscillators and Bayesian Decision Makers



Disbelievers have significantly lower expectations from autonomy compared to Oscillators and Bayesian Decision Makers



** $p < 0.01$; * $p < 0.05$; † $p < 0.1$
BDM=Bayesian Decision Maker

Conclusion and Future Work

- Knowing the type of trust dynamics of an individual could influence whether a machine partner with a dynamic trust model is a feasible solution for that individual
- We assume that the human behaves according to a reverse psychology model. The study should be expanded to include more advanced models of human behavior
- Inverse Reinforcement Learning techniques can be used to learn personalized reward function weights to further improve trust estimation and team performance

References

1. Y. Guo and X. J. Yang, "Modeling and predicting trust dynamics in human-robot teaming: A Bayesian inference approach," International Journal of Social Robotics, 12 2021.
2. 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.
3. A. Xu and 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.

Acknowledgement

This work was supported in part by the Air Force Office of Scientific Research (Grant #FA9550-20-1-0406).