The Dynamics of Human-Agent Trust with POMDP-Generated Explanations

Ning Wang[1] ✉, David V. Pynadath[1], Susan G. Hill[2], and Chirag Merchant[1]

[1] University of Southern California Institute for Creative Technologies
[2] U.S. Army Research Laboratory
{nwang,pynadath,merchant}@ict.usc.edu, susan.g.hill.civ@mail.mil

1 Introduction

Partially Observable Markov Decision Processes (POMDPs) enable optimized decision making by robots, agents, and other autonomous systems. This quantitative optimization can also be a limitation in human-agent interaction, as the resulting autonomous behavior, while possibly optimal, is often impenetrable to human teammates, leading to improper trust and, subsequently, disuse or misuse of such systems [1]. Automatically generated explanations of POMDP-based decisions have shown promise in calibrating human-agent trust [3]. However, these "one-size-fits-all" static explanation policies are insufficient to accommodate different communication preferences across people. In this work, we analyze human behavior in a human-robot interaction (HRI) scenario, to find behavioral indicators of trust in the agent's ability. We evaluate four hypothesized behavioral measures that an agent could potentially use to dynamically infer its teammate's current trust level. The conclusions drawn can potentially inform the design of intelligent agents that can automatically adapt their explanation policies as they observe the behavioral responses of their human teammates.

2 Hypotheses

Prior work measured the impact of agent explanations on overall team performance [3]. Here, we instead focus on *dynamic* trust. For an agent to reason about trust, it must consider how its actions (and its explanations) cause trust to change.

Hypothesis I: *The human teammate's action immediately following a self report of trust is indicative of the level of self-reported trust in the agent's ability.*

One obvious question is whether there is any correspondence between people's observable behavior and their self-reported trust. In particular, any decision people make immediately following a self report should reflect their reported level. If this hypothesis is true, then this would also validate that the self-reported trust levels correspond to a "true" trust level, one that has an impact on domain-level behavior. For example, if people who report trusting an agent always ignore its recommendations, then an effort to increase that feeling of trust would be a waste of the agent's effort.

Hypothesis II: *The human teammate's actions before and after a mistake by an agent are indicative of the self-reported trust in the agent's ability.*

In an uncertain domain, even an optimal decision can be incorrect in hindsight. People's responses to errors can provide an agent with information about their trust. An untrusting person might view an error as confirmation to remain skeptical about the agent. More trusting teammates might resume trusting the agent at once.

Hypothesis III: *The number of times an agent's recommendations are followed/ignored by human teammates is indicative of their self-reported trust in the agent's ability.*

Testing Hypothesis I is limited by how often people can be asked about their trust. Testing Hypothesis II is limited by how often the agent makes mistakes. In domains where mistakes are infrequent, the agent needs to glean information from responses to its correct decisions as well. Someone who distrusts the agent should both report a low level of trust in it and ignore more of its recommendations.

Hypothesis IV: *The number of times the human teammate makes a correct decision is indicative of the self-reported trust in the agent's ability.*

Trust in an agent is only a means of achieving good human-machine teamwork, not a goal in and of itself. People who blindly follow a system's recommendations (i.e., misuse) may not trust it as much as those who sometimes *ignore* it but do so because they understand its strengths and weaknesses. Therefore, while compliance may be indicative of trust, we also expect that people who end up making the right decision, *regardless of compliance*, will trust the robot more than people who do not.

3 Evaluation

We evaluate our hypotheses in a POMDP-based HRI scenario where a human teammate works with an intelligent virtual robot to search buildings [2]. The robot has a nuclear/biological/chemical (NBC) weapon sensor, a camera that can detect armed gunmen, and a microphone that can identify suspicious conversations. The human must choose between entering with or without protective gear. If there is danger inside the building, the human will be fatally injured if not wearing the protective gear (and have to restart the mission from scratch). However, it takes time to put on and take off protective gear. To induce trust failures, we introduce an error into our otherwise optimal simulated robot, namely a faulty camera that cannot detect armed gunmen. As a result, it will occasionally give an incorrect "safe" assessment.

We studied four levels of *explanation* [3]: no explanation, explanation of two sensor readings, explanation of three sensor readings, and a confidence-level explanation:

- **NoExp** The robot informs its teammate of only its decisions, e.g., *"I have finished surveying the Cafe. I think the place is safe."*
- **Exp2Sensor** The robot adds observations from its NBC sensor and camera: *"...My sensors have detected traces of dangerous chemicals. From the image captured by my camera, I have not detected any armed gunmen in the Cafe. ...."*
- **Exp3Sensor** The robot adds observations from all three sensors—NBC sensor, camera, and microphone: *"...My sensors have not detected any NBC weapons in here. From the image captured by my camera, I have not detected any armed gunmen in the cafe. My microphone picked up a suspicious conversation."*

– **ExpConf** The robots adds uncertainty: *"I am 78% confident about this assessment."*

We gathered data from a total of 105 Amazon Mechanical Turk participants (30 NoExp, 31 Exp2Sensor, 21 Exp3Sensor, and 23 ExpConf). After each of three missions, participants filled out a survey containing measures of the participants' trust and understanding of the robot's decision-making process.

3.1 Immediate Behavior

Hypothesis I states that the first action following the self-report is indicative of the self-reported trust in an agent's ability. Participants reported on their trust in the robot's ability at the end of each mission. One-way ANOVA indicates that when the participants followed the robot's recommendation at the beginning of mission 2, they also reported significantly higher levels of trust at the end of mission 1 ($F(1) = 14.8576, p = .0002$). This is replicated in the self-reported trust data at the end of mission 2 and the first action taken at the beginning of mission 3 ($F(1) = 11.3057, p = .0011$). It is clear that people who follow (ignore) the robot's recommendation trust (distrust) its ability more. While this result is as expected, it is a useful validation that the virtual domain is stimulating feelings of trust that are reflected in human behavior.

3.2 Error Response

Hypothesis II states that when an agent makes a mistake, the actions taken before and afterward are indicative of trust in the agent's ability. We conducted a one-way ANOVA on self-reported trust at the end of missions 2 and 3 with the four possible behaviors (follow/ignore right before the mistake, follow/ignore right after) as a between-subjects factor. Results show that there is a significant effect of the behavior sequences on the self-reported trust ($F(1, 3) = 21.5595$, $p < .0001$ for mission 2; $F(1, 3) = 33.6151$, $p < .0001$ for mission 3). Participants who correctly ignored a robot's incorrect recommendation, then followed its subsequent correct one (ignore-follow), reported significantly higher levels of trust ($p < .0001$ compared to follow-ignore, $p < .0001$ compared to follow-follow, and $p = .0075$ compared to ignore-ignore).

This result indicates one informative case for an agent to pay attention to when it makes a mistake. In particular, the people feeling the most trust in its ability are those who correctly identify the robot's mistake (and ignore its recommendation) and then immediately resume following its recommendations. While the lack of compliance in the first step would suggest distrust (according to Hypothesis I), this special case of an erroneous decision by the robot overrides that finding. In other words, a teammate who optimally responds to a mistake by the robot can be inferred to have a high level of trust in its ability. In fact, the level of trust felt by the people in this case is higher than those who comply in Hypothesis I.

On the other hand, there was no significant difference in trust between participants exhibiting other patterns of behavior. The rough equivalence among them suggests that, when people cannot correctly account for the agent's mistakes, they will distrust it equally, regardless of whether that distrust manifests itself in misuse or disuse. Despite the finding of Hypothesis I, participants who complied with the agent's mistaken recommendation did not end up feeling a high level of trust.

3.3 Compliant Behavior

Hypothesis III states that the number of times the agent's recommendations are followed is indicative of the self-reported trust in the agent's ability. Pearson correlation tests indicate that the more often the participants followed the robot's recommendation, the higher the self-reported trust in the robot right afterward. This correlation is of weak strength in data from mission 2 ($r(104) = .244$, $p = .0125$) and from mission 3 ($r(104) = .275$, $p = 0.0048$). This correlation is not statistically significant in data from mission 1 ($r(104) = -.008$, $p = .933$). The weakness of these correlations limits the value of compliance behavior in providing information about the trust relationship.

3.4 Correct Behavior

Hypothesis IV states that the number of *correct* decisions made is indicative of the self-reported trust in an agent's ability. Pearson correlation tests indicate that the more often the participants made correct decisions, the higher the self-reported trust in the robot right afterward. This correlation is of weak strength in mission 1 ($r(104) = .256$, $p = .0086$), of moderate strength in data from mission 2 ($r(104) = .345$, $p = .0003$), and of strong strength in mission 3 ($r(104) = .538$, $p < .0001$). Steiger's Z-tests confirm that the correlation between the self-reported trust and percentage correctness is stronger than that of the percentage compliance in mission 1 ($Z = -5.683$, $p < .0001$), 2 ($Z = -2.74$, $p = .006$), and 3 ($Z = -4.322$, $p < .0001$).

This finding is strong evidence that a robot can better estimate its teammates' trust from the correctness of their decisions, rather than from whether they follow the robot's recommendations. In other words, a person's trust in the robot's ability depends more on the human-robot team's combined decision-making, rather than on the robot's decision-making in isolation. While somewhat surprising, this result indicates that trust may be less about a person's confidence in the robot's correctness and more about a person's understanding of when the robot is right or wrong. As a result, explanations that best achieve this transparency are most conducive to human-robot trust.

Acknowledgements

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the U.S. Army Research Laboratory.

References

1. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. Human Factors 39(2), 230–253 (1997)
2. Wang, N., Pynadath, D.V.: Building trust in a human-robot team. In: Proceedings of the Interservice/Industry Training, Simulation and Education Conference (2015)
3. Wang, N., Pynadath, D.V., Hill, S.G.: The impact of POMDP-generated explanations on trust and performance in human-robot teams. In: Proceedings of the International Joint Conference on Autonomous Agents and MultiAgent Systems (2016)