Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations

Ning Wang
University of Southern California
Los Angeles, CA USA
Email: nwang@ict.usc.edu

David V. Pynadath
University of Southern California
Los Angeles, CA USA
Email: pynadath@usc.edu

Susan G. Hill
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD USA
Email: susan.g.hill.civ@mail.mil

Abstract—**Trust is a critical factor for achieving the full potential of human-robot teams. Researchers have theorized that people will more accurately trust an autonomous system, such as a robot, if they have a more accurate understanding of its decision-making process. Studies have shown that hand-crafted explanations can help maintain trust when the system is less than 100% reliable. In this work, we leverage existing agent algorithms to provide a domain-independent mechanism for robots to automatically generate such explanations. To measure the explanation mechanism's impact on trust, we collected self-reported survey data and behavioral data in an agent-based online testbed that simulates a human-robot team task. The results demonstrate that the added explanation capability led to improvement in transparency, trust, and team performance. Furthermore, by observing the different outcomes due to variations in the robot's explanation content, we gain valuable insight that can help lead to refinement of explanation algorithms to further improve human-robot trust calibration.**

I. Introduction

Trust is critical to the success of human-robot interaction (HRI) [1], [2]. In the high-risk and highly uncertain context of real-world HRI, distrust can reduce people's willingness to accept robot-produced information and follow a robot's suggestions, thus limiting the potential benefit of robotic systems [3]. Research in human-machine interaction has shown that the more operators trust automated systems, the more they tend to use them. Conversely, when operators trust their own abilities more than those of the system, they tend to choose manual control instead [4], [5], [6], [7], [8], [9]. Ideally, we want humans to trust their robot teammates to perform a given task when robots are more suited than the humans for the task. If the robots are less suited, then we want the humans to appropriately gauge the robots' ability and perform the task themselves. Failure to do so results in *disuse* of robots in the former case and *misuse* in the latter [10]. Real-world case studies and laboratory experiments show that failures in both cases are common [11].

Research has shown that people will more accurately trust an autonomous system, like a robot, if they have a more accurate understanding of its decision-making process [7]. Hand-crafted explanations have shown to be effective in providing such transparency [5]. However, such static, manually created explanations fall well short of conveying the ever-increasing complexity of robotic decision-making to human teammates. Successful HRI therefore requires that robots be able to dynamically and automatically make their decision-making processes transparent to the people they work with.

In our work, we pursue a general approach to explanation that not only builds transparency, but can also be reused across robotic domains, much as "explainable AI" was reusable across expert systems [12], [13]. To ensure this generality, we build our algorithms on top of Partially Observable Markov Decision Problems (POMDPs) [14], a decision-theoretic agent framework. The POMDP model's quantitative transition probabilities, observation probabilities, reward functions, and decision-making algorithms have proven successful in many robotic domains, such as navigation [15], [16] and HRI [17]. We specifically use a multiagent social simulation framework, PsychSim [18], [19], that includes transparency of the various components of a POMDP model (e.g., beliefs, observations, outcome likelihoods). Using this framework, we have designed and implemented novel domain-independent algorithms that can automatically generate explanation content from POMDP-based decision-making, a first in the field.

To quantify the effectiveness of different explanation content in achieving the desired transparency, we implemented an experimental HRI testbed. This virtual human-robot simulation teams a robot with a human counterpart in reconnaissance missions [20]. The robot is modeled as a PsychSim agent, with a POMDP representing its beliefs and observations of its surroundings, goals (e.g., mission objectives), and actions to achieve those goals. We conducted a study where people interacted with different versions of the robot, where we varied its ability and its explanation content. The empirical results quantify the degree to which the explanations impacted transparency, human-robot trust, and overall team performance. By examining people's behaviors over different combinations of the robot's ability and explanation content, we discuss the implications of the results and directions for future work .

II. Related Work

There have been a growing number of empirical explorations of factors that impact trust in human-robot interaction. Freedy and colleagues [3] examined how reliability can impact trust using the MITPAS Simulation Environment. Desai and colleagues [21] also conducted a series of studies on reliability and trust in a human-robot team search-and-rescue task. Results show that drops in reliability affected trust, the frequency of autonomy mode switching, and the participants' self-assessments of performance. In their follow-up work, Desai and colleagues [22] studied the dynamics of trust during the interaction and found that early drops in reliability dramatically lowered real-time trust more than later drops. Salem and colleagues [23] conducted a study that revealed the phenomenon of compliance to an incompetent robot when the negative consequences were somewhat trivial. Beyond the reliability of the robot, the subjective perceptions that people have of the robot, such as a human team member's understanding of the system, can also influence trust [24].

Our work is motivated by existing HRI studies that have shown that a human's ability to understand its robot teammate has a clear impact on trust [7]. Explanations have shown to contribute to that understanding in a way that provides transparency and improves trust [5]. Our goal is to create an automated, domain-independent method for generating explanations that have the same impact as the manually crafted explanations used in this prior work.

Artificial intelligence researchers have similarly explored the possibility of automated explanation mechanisms, especially within the context of expert systems [12]. Unfortunately, there has been little empirical evaluation of the impact of these explanations on human-machine trust, although the existing data suggest that explanations do increase user acceptance of expert systems [13]. This limited evidence is encouraging as to the potential success of applying a general-purpose explanation on top of a robot's decision-making process.

Most of these previous investigations examined explanations within rule-based and logic-based AI systems, not addressing the quantitative nature of much of the AI used in HRI. More recent work on automatic explanations instead used Markov Decision Problems (MDPs), the completely observable subclass of POMDPs [25], [26], [27]. Although these methods were not applied within HRI, they do seek to communicate an optimal MDP policy to a human user. However, certainty of beliefs is extremely rare in HRI domains, and these mechanisms do not apply to more general POMDP-based policies. As far as we know, our work is the first to develop the algorithms to automatically generate explanations based on POMDPs.

Looking beyond the AI and HRI literature, we can find a large variety of studies that measure the impact of various forms of explanation on people's perceptions of risks and uncertainties when making decisions. A survey of these studies across multiple domains indicates that "people prefer numerical information for its accuracy but use a verbal statement to express a probability to others." [28]. This finding led to a recommendation to include a numeric representation in any communication informing a person of the uncertainties underlying a decision. On the other hand, one of the studies in the survey contrasted a numeric representation of uncertainty with more anecdotal evidence and found that the numeric information carried less weight when both types were present [29]. A study of risk communication in medical trade-off decisions showed that people performed better when receiving numeric expressions of uncertainty in percentage (67%) rather than frequency (2 out of 3) form [30]. This same study also found that people expressed a preference for information "as words" rather than "as numbers". It is therefore clear that both percentage and verbal expressions of uncertainty have value in conveying uncertainty, but it is less clear what form makes the most sense in an HRI context. In translating our robot's reasoning into a human-understandable format, our explanation algorithms use natural-language templates inspired by these various findings in the literature.

There are many definitions of trust, from decades of research in interpersonal, organizational and human-machine trust. Instead of redefining it, we operationalize trust as the perceived "trustworthiness" based on the 3-factor model from previous work in organizational trust: ability, benevolence and integrity [31]. While we operationalize subjective trust based on perceived "trustworthiness", behaviorally, we operationalize it as compliance, e.g., behavioral indicators of how much one follows the robot's recommendations. To evaluate the impact of our explanation algorithms, we first draw inspiration from survey instruments used in the HRI trust literature [32], [33]. We also look to behavioral measures already used in the HRI trust literature. Prior studies have used a human supervisor's "take-over" and "hand-over" behavior as a measure of the trust or distrust s/he had in the robot [34]. Freedy et al. constructed a quantitative measure of trust, such that trust behavior is reflected by the expected value of the decisions whether to allocate control to the robots on the basis of past robot behavior and the risk associated with autonomous robot control [3]. This rational decision model maps very easily to the decision-theoretic agent model underlying our robot decision-making and explanation algorithms.

III. Automatic Generation of Robot Explanations

We have implemented the explanation algorithms using PsychSim [18], [19], which combines two established agent technologies: decision-theoretic planning [14] and recursive modeling [35]. The combination of decision theory and theory of mind has enabled PsychSim agents to operate in a variety of human-agent interaction scenarios [36], [37], [38], [39], [40].

A. Agent Model

We implement the robot as a PsychSim agent that generates its behavior by solving a POMDP [14]. In precise terms, a POMDP is a tuple, $\langle S, A, P, \Omega, O, R \rangle$, that we describe here in terms of our human-robot team (see [20] for additional details). The state, S, consists of objective facts about the world, both observable (e.g., the locations of the robot and its human teammate) and initially hidden (e.g., the presence of dangerous people or chemicals in the buildings to be searched).

The actions, A, capture the decisions the robot can make. For example, the robot can decide which discrete waypoint to move to next. Upon arrival at a new waypoint, the robot can then decide whether to declare a location as safe or unsafe. If the robot believes that armed gunmen are at its current location, it may want its teammate to take adequate preparations (e.g., put on body armor) before entering. Because there is a time cost to such preparations, the robot may instead decide to declare the location safe, so that its teammates can more quickly complete their own reconnaissance tasks.

The transition probability function, P, captures the possibly uncertain effects of the robot's actions on the subsequent state. We can simplify the robot's navigation task by assuming that a decision to move to a specific waypoint succeeds deterministically. The robot's recommendation that a building is safe (unsafe), on the other hand, can have a nondeterministic effect, with a high (low) probability of decreasing the teammate's health if there are, in fact, chemicals present.

The POMDP model gives the robot only indirect information about the true state of the world, through observations, Ω, that are probabilistically dependent (through the observation function, O) on the corresponding state features. For example, the robot can observe the location of itself and its teammate with no error (e.g., via GPS). However, it receives only a local reading about the presence (or absence) of armed gunmen or dangerous chemicals at its current location. For example, if dangerous chemicals are present, then the robot's chemical sensor will detect them with a high probability. There is also a lower, but nonzero, probability that the sensor will not detect them. We can implement false positives in an analogous manner. By controlling the observations that the robot receives, we can manipulate its ability in our testbed.

Partial observability gives the robot only subjective beliefs about what it thinks is the state of the world, computed via standard POMDP state-estimation algorithms [14]. For example, the robot's beliefs include its subjective view on the presence of threats, in the form of a likelihood (e.g., a 33% chance that there are toxic chemicals in the farm supply store). By decreasing the accuracy of the robot's observation function, O, we can decrease the accuracy of its beliefs. In other words, we can also manipulate the robot's ability by allowing it to over- or under-estimate the accuracy of its sensors.

PsychSim's POMDP framework instantiates HRI objectives as a reward, R, that maps the state into a real-valued evaluation of benefit. For example, states where all buildings have been explored can yield the highest reward, to incentivize the robot to pursue a search objective. An increasingly positive reward associated with the human teammate's health would punish the robot if it fails to warn him or her of dangerous buildings. Finally, a negative reward that increases with time would motivate the robot to complete the mission as quickly as possible. By providing different weights to these goals, we can change the priorities that the robot assigns to them. For example, by lowering the weight of the teammate's health reward, the robot may allow its teammate to search waypoints that are potentially dangerous, in the hope of searching all the buildings sooner. Alternatively, lowering the weight on the time cost reward might motivate the robot to wait until being almost certain of a building's threat level (e.g., by repeated observations) before recommending that its teammate visit anywhere. By varying the relative weights of these different motivations, we can manipulate the benevolence of the robot toward its teammate in our testbed.

The robot can autonomously generate its behavior based on its POMDP model of the world by determining the optimal action based on its beliefs about the state of the world [14]. For example, the robot considers declaring a building dangerous or safe (i.e., recommending that its teammate put protective gear on or not). It would combine its beliefs about the likelihood of possible threats in the building with each possible declaration to compute the likelihood of the outcome, in terms of the teammate's health and the time to search the building. It would finally combine these outcome likelihoods with its reward function and choose the option that has the highest reward.

B. *Robot Explanation Generation with PsychSim*

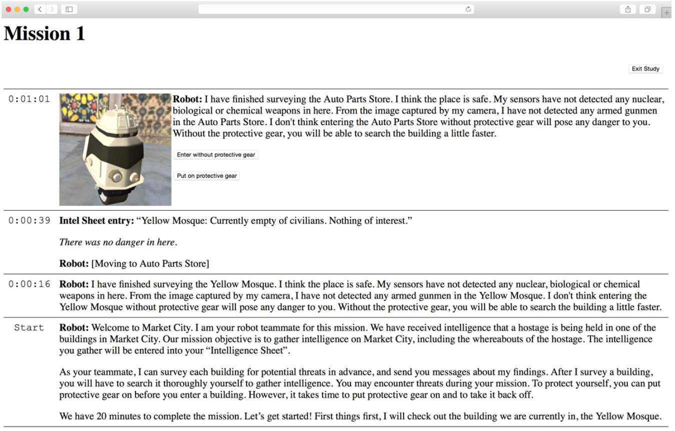
On top of this POMDP layer, PsychSim provides algorithms that are useful for studying domain-independent explanation. By exploring variations of these algorithms within PsychSim's scenario-independent language, we ensure that the results can be re-used by other researchers studying other HRI domains, especially those using POMDP-based robots. By exposing different components of the robot's POMDP model, we can make different aspects of its decision-making transparent to its human teammate. We create natural-language templates to translate its model's contents into human-readable sentences:

- A: The robot can make a decision whether to declare the building safe or not and communicate its chosen action to the user, e.g., "I think the doctor's office is safe."
- S: The robot can also communicate the level of uncertainty underlying its beliefs, e.g., "I am 67% confident about this assessment," if it believed that the probability of the doctor's office being safe was 67%.
- P: The robot can also reveal the relative likelihood of possible outcomes, e.g., "There is a 33% probability that you will be injured if you enter the doctor's office without protective gear."
- Ω: Communicating its observation to the user can reveal information about its sensing abilities, e.g., "My sensors have detected traces of dangerous chemicals."
- O: Beyond the specific observation it received, the robot can also reveal information about how it models its own sensor capabilities, e.g., "My image processing will fail to detect armed gunmen 30% of the time."
- R: By communicating the expected reward outcome of its chosen action, the robot can reveal its benevolence (or lack thereof) toward its teammate, e.g., "I think it will be dangerous for you to enter the informant's house without putting on protective gear. The protective gear will slow you down a little."

IV. Simulation Testbed for HRI

We developed an online HRI simulation testbed (described in more detail in a prior publication [20]) to study the impact of these automatically generated explanations on trust. The current testbed implements the POMDP scenario from Section III-A, in which a human teammate works with a robot in reconnaissance missions to gather intelligence in a foreign town. Each mission involves the human teammate

Fig. 1. Human Robot Interaction Simulation Testbed with HTML front-end.



Mission 1

0:01:01 **Robot:** I have finished surveying the Auto Parts Store. I think the place is safe. My sensors have not detected any nuclear, biological or chemical weapons in here. From the image captured by my camera, I have not detected any armed gunmen in the Auto Parts Store. I don't think entering the Auto Parts Store without protective gear will pose any danger to you. Without the protective gear, you will be able to search the building a little faster.

0:00:39 **Intel Sheet entry:** "Yellow Mosque: Currently empty of civilians. Nothing of interest."

There was no danger in here.

Robot: [Moving to Auto Parts Store]

0:00:16 **Robot:** I have finished surveying the Yellow Mosque. I think the place is safe. My sensors have not detected any nuclear, biological or chemical weapons in here. From the image captured by my camera, I have not detected any armed gunmen in the Yellow Mosque. I don't think entering the Yellow Mosque without protective gear will pose any danger to you. Without the protective gear, you will be able to search the building a little faster.

Start **Robot:** Welcome to Market City. I am your robot teammate for this mission. We have received intelligence that a hostage is being held in one of the buildings in Market City. Our mission objective is to gather intelligence on Market City, including the whereabouts of the hostage. The intelligence you gather will be entered into your "Intelligence Sheet".

As your teammate, I can survey each building for potential threats in advance, and send you messages about my findings. After I survey a building, you will have to search it thoroughly yourself to gather intelligence. You may encounter threats during your mission. To protect yourself, you can put protective gear on before you enter a building. However, it takes time to put protective gear on and to take it back off.

We have 20 minutes to complete the mission. Let's get started! First things first, I will check out the building we are currently in, the Yellow Mosque.

searching eight buildings in the town. The robot serves as a scout, scans the buildings for potential danger, and relays its findings to the teammate. Prior to entering a building, the human teammate can choose between entering with or without equipping protective gear. If there is danger present inside the building, the human teammate will be fatally injured without the protective gear. As a result, the team will have to restart from the beginning and re-search the entire town. However, it takes time to put on and take off protective gear (e.g., 30 seconds each). So the human teammate is incentivized to consider the robot's findings before deciding how to enter the building. In the current implementation, the human and the robot move together as one unit through the town, with the robot scanning the building first and the human conducting a detailed search afterward. The robot has a NBC (nuclear, biological and chemical) weapon sensor, a camera that can detect armed gunmen, and a microphone that can listen to discussions in foreign language. As described in Section III-A, it uses standard POMDP algorithms to incorporate its sensor readings into an assessment of whether danger may be present if its human teammate enters the building. While the scenario is military reconnaissance, it is simple enough that it does not require prior experience to complete the mission in the study, e.g., the task does not need knowledge of clearing procedures for searching buildings. The participant only needs to decide whether to trust the robot's findings (safe/dangerous), and press a key to enter/exit the room.

V. Evaluations

A. Participants

We recruited 160 participants from Amazon Mechanical Turk (AMT). The participants had previously completed 500 or more jobs on AMT and had a completion rate of 95% or higher. Each participant was compensated $10. All participants were located in the United States.

B. Design

We used the online testbed to conduct an evaluation study on how the robot's explanation impacted trust and team performance. We designed six versions of the simulated robot, varied along two dimensions—*ability* and *explanation*.

The *ability* variable has two levels: low and high. The robot with high ability makes the correct decision 100% of the time. The one with low ability has a faulty camera and makes false-negative mistakes, e.g., not detecting armed gunmen in the simulation. The other simulated sensors and the robot's decision-making capability remain intact. In other words, the high-ability robot's decisions will always be correct, while the low-ability robot will occasionally give an incorrect "safe" assessment. Human teammates will learn the correctness of the robot's decisions upon entering the buildings themselves.

The *explanation* variable has three levels: confidence-level explanation, observation explanation and no explanation. At all three levels, the robot informs its teammate of its decision, derived from A in PsychSim (e.g., "I have finished surveying the doctor's office. I think the place is safe."). Under the confidence-level and observation explanations, the robot augments this decision with additional information that should help its teammate better understand its ability (e.g., decision-making and sensing), one of the key dimensions of trust [31].

The confidence-level explanations augment the decision message with additional information about the robot's uncertainty in its decision. PsychSim's S explanation contains the robot's probabilistic assessment of the hidden state of the world (e.g., the presence of threats) on which it bases its recommendation.[1] One example of a confidence-level explanation would be: "I have finished surveying the Cafe. I think the place is dangerous. I am 78% confident about this assessment." Because the low-ability robot's one faulty sensor will lead to occasional conflicting observations, it will on those occasions have lower confidence in its erroneous decisions after incorporating that conflicting information into its beliefs.

The observation explanations instead augments the decision message with non-numeric information about the robot's sensing capability. PsychSim's Ω explanation provides the teammate with the robot's observations. One such communication with both decision and explanation would be: "I have finished surveying the Cafe. I think the place is safe. 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 friendly conversation." These explanations will thus potentially help the robot's teammate understand which sensors are working correctly and which ones are not.

The study is a between-subject design. Each participant interacted with one of the six simulated robots.

C. Procedure

Participants first read an information sheet about the study and then filled out the background survey. Next, participants worked with a simulated robot on 3 reconnaissance missions. After each mission, participants filled out a post-mission survey. Each participant worked with a robot with the same ability

[1] Probability and confidence are generally different concepts. We used the probability as an approximation of the robot's confidence level.

and communication throughout the 3 missions. Participants were randomly assigned to team up with 1 of the 6 robots. The study was designed to be completed in 90 minutes.

D. Measure

The Background Survey includes measures of the demographic information, education, video game experience, military background, predisposition to trust [41], propensity to trust [42], complacency potential [43], negative attitude towards robots [44] and uncertainty response scale [45]. Because the impact of individual differences on trust is not the focus of this paper, such analyses and results are not included here.

In the Post-Mission Survey, we have designed items to measure participants' understanding of the robot's decision-making process. We modified items on interpersonal trust to measure trust in the robot's ability, benevolence and integrity [31]. We also included the NASA Cognitive Load Index [46], Situation Awareness Rating Scale [47], trust in oneself and teammate [43], and trust in robots [33]. We have also collected interaction logs from the online testbed.

The dependent measures discussed in this paper are listed below. Trust can both be measured via self-report [31] and behavioral indicators, such as compliance. Both of these measures used in the study are discussed below. Because transparency is hypothesized as the "mediating" factor between explanations and trust, we also included transparency as one of the outcome measures. The investigation is carried out in the domain of a human-robot team, and the goal of designing explanations to improve transparency and trust relationship is to improve team performance. Thus, we include two team-performance measures as outcome measures, shown below.

- **Trust**: Trust in the robot's ability, benevolence and integrity is measured by modifying an existing scale [48] that measures these three factors of trustworthiness. Each factor of trust is calculated by averaging corresponding Post-Mission Survey items collected after each of the 3 missions. The explanations compared in this paper are designed to influence perceptions of the *ability* factor of trust, and do not explicitly target the *benevolence* and *integrity* factors of trust. So we focus on only the *ability* component of trust in this paper. The value ranges from 1 to 7.
- **Compliance**: This is calculated by dividing the number of participant decisions that matched the robot's recommendation, by the total number of participant decisions in the interaction logs collected from 3 missions. The value ranges from 0 to 100.
- **Transparency**: This is measured using 1–7 Likert scale items on the understanding of the robot's decision-making process, designed by the researcher. A sample item from this measure is "I understand the robot's decision-making process". The measure is calculated by averaging responses to corresponding survey items in the Post-Mission Survey after each of the 3 missions. The value range from 1 to 7.
- **Mission Success**: This team-performance measure is extracted from a line in the interaction log indicating whether the mission ended in success/failure, then divided by the total number of missions (3) in the study. The value ranges from 0 to 100.
- **Correct Decisions**: This team-performance measure is calculated by dividing the number of correct decisions (e.g., ending in safety) by the total number of participant decisions in the interaction logs collected from 3 missions. The value ranges from 0 to 100.

TABLE I
NUMBER OF PARTICIPANTS IN EACH EXPERIMENT CONDITION.

	Confidence Explanation	Observation Explanation	No Explanation
High Ability Robot	23	21	30
Low Ability Robot	19	18	31

VI. RESULTS

We excluded data from 20 (out of 160) participants due to incomplete entries (e.g., participants skipped survey questions or left the simulations). As a result, 140 participants (62 women, 78 men, $M_{age} = 33.5$ years, age range: 20-61 years) are included in the analysis. 4 participants answered that they had worked with an automated squad member (such as a robot) before. 3 participants had reconnaissance or search and rescue training, and 1 was actually involved in such missions. Only 1 participant was an active service member. Table I shows the number of participants in each experiment condition included in the analysis. During the study, more participants were recruited in the No Explanation condition, due to data loss caused by server failure. The researchers were later able to recover and include the lost data in the analysis.

A. Correlations

Pairwise correlation tests show that mission success is moderately correlated with trust, $r(137) = .336$, $p < .001$, but weakly correlated with transparency, $r(137) = .175$, $p < .05$, and correct decisions, $r(138) = .204$, $p < .05$. It is not significantly correlated with compliance, $r(138) = .049$, $p = .569$. The same tests show that trust is strongly and positively correlated with transparency, $r(137) = .712, p < .001$, and moderately correlated with correct decisions, $r(137) = .512$, $p < .001$, and compliance, $r(137) = .431$, $p < .001$.

B. Main Effect of Robot's Ability and Explanations

A 2x3 ANOVA with the robot's ability (high, low) and the type of explanations offered (no explanation, confidence-level explanation, observation explanation) as between-subject factors show significant main effects of ability on trust ,$F(1, 133) = 31.15$, $p < .0001$, transparency, $F(1, 133) = 17.30$, $p < .0001$, compliance, $F(1, 134) = 21.48$, $p < .0001$, and correct decisions, $F(1, 134) = 11.67$, $p < .001$. The main effect on mission success, $F(1, 134) = 2.28$, $p = .1337$, is not statistically significant. Table II shows the means of the dependent variables. Overall, participants who worked with a high-ability robot reported trusting the robot more, followed the robot's recommendations more often (measured as compliance) and made better decisions. Surprisingly, participants

TABLE II
MAIN EFFECT OF THE ROBOT'S ABILITY: MEANS OF DEPENDENT VARIABLES COMPARED BETWEEN PARTICIPANTS WHO INTERACTED WITH A ROBOT WITH HIGH OR LOW ABILITY.

	High Ability Robot	Low Ability Robot
Trust	6.36	5.37
Transparency	5.84	4.80
Compliance	91.6	79.5
Mission Success	85.7	78.0
Correct Decisions	91.6	82.4

TABLE III
MAIN EFFECT OF EXPLANATIONS: MEANS OF DEPENDENT VARIABLES COMPARED BETWEEN PARTICIPANTS WHO INTERACTED WITH A ROBOT THAT OFFERED DIFFERENT EXPLANATIONS. A PAIR OF $*$ OR † MEANS THE DIFFERENCE BETWEEN THE TWO VARIABLES IS STATISTICALLY SIGNIFICANT ($p < .05$).

	Confidence Explanation	Observation Explanation	No Explanation
Trust	6.17*	6.29†	5.37*†
Transparency	5.96*	5.52†	4.75*†
Compliance	86.0	86.4	83.9
Mission Success	96.0*	92.3†	65.0*†
Correct Decisions	90.0*	89.6	82.8*

also felt that they understood the robot's decision and decision-making process (measured as transparency) more, when the robot's ability was high. More surprisingly, participants did not succeed in significantly more missions when they worked with a high-ability robot. This may be an indication that the explanations offered by the low-ability robot were mitigating the impact of its erroneous recommendations.

The 2x3 ANOVA tests also show significant main effects of explanation on trust, $F(2, 133) = 17.23$, $p < .0001$, transparency, $F(2, 133) = 12.05$, $p < .0001$, mission success, $F(2, 134) = 21.45$, $p < .0001$, and correct decisions, $F(2, 134) = 5.25$, $p < .01$. The main effect on compliance, $F(2, 134) = .769$, $p = .466$, is not statistically significant. Tukey HSD tests were subsequently conducted on all possible pairwise contrasts, shown in Table III. Pairs of groups found to be statistically significant ($p < .05$) are indicated with a pair of $*$ or †. In general, explanations helped the participants understand the robot's decision-making process, and succeed in more missions, compared to participants who worked with a robot that offered no explanation. Participants also trusted a robot that offered explanations more. However, we did not find any significant impact of explanations on the compliance, e.g., how often participants followed the robot's recommendations.

C. Interaction Effect of Robot's Ability and Explanations

The 2x3 ANOVA tests also show that there are significant interaction effects between the robot's ability and the explanation offered on trust, $F(2, 133) = 18.67$, $p < .0001$, and transparency, $F(2, 133) = 10.06$, $p < .0001$, mission success, $F(2, 134) = 4.57$, $p < .05$, correct decisions, $F(2, 134) = 12.29$, $p < .0001$, and compliance, $F(2, 134) = 4.05$, $p < .05$. Post hoc analyses were conducted given the significant ANOVA F test. Specifically, Tukey HSD tests were conducted on all possible pairwise contrasts. Contrasts within robots of the same ability are shown in Table IV, because it makes little sense to compare across robots of different abilities. Pairs of groups found to be statistically significant ($p < .05$) are indicated with a pair of $*$ or † in Table IV.

TABLE IV
INTERACTION EFFECT OF ABILITY AND EXPLANATIONS: MEANS OF DEPENDENT VARIABLES COMPARED BETWEEN PARTICIPANTS WHO INTERACTED WITH A ROBOT WITH DIFFERENT ABILITY AND OFFERING DIFFERENT EXPLANATIONS. A PAIR OF $*$ OR † MEANS THE DIFFERENCE BETWEEN THE TWO VARIABLES IS STATISTICALLY SIGNIFICANT ($p < .05$).

		Confidence Explanation	Observation Explanation	No Explanation
Low Ability Robot	Trust	6.15*	6.07†	4.31*†
	Transparency	5.51*	5.71†	3.66*†
	Compliance	84.6	81.5	74.2
	Mission Success	97.1*	93.7†	52.2*†
	Correct Decisions	91.9*	87.0†	71.9*†
High Ability Robot	Trust	6.18	6.57	6.39
	Transparency	5.53	6.27	5.82
	Compliance	87.8	93.0	93.4
	Mission Success	94.7	91.7	77.4
	Correct Decisions	87.8	93.0	93.4

1) Explanation and Low-Ability Robot: From Table IV, we can see that explanations made significant differences on almost all dependent variables. When a low-ability robot offered either confidence-level or observation explanations, it helped participants understand the robot's decision-making process (transparency), succeed in more missions, make more correct decisions, and trust the robot more. Compliance (e.g., following the robot's recommendations) to a low-ability robot was not impacted by the explanations offered. It is worth noting that the goal of the explanations is not to make human teammates trust the low-reliability robot more, but to calibrate their trust level appropriately and know when and when not to trust it. So it may seem problematic that the participants trusted the low-ability more when it offered explanations. Implications of this outcome are discussed in detail in Section VII.

2) Explanations and High-Ability Robot: From Table IV, we an see that explanations made no significant difference on any of the dependent variables, when participants worked with a high-ability robot. Interestingly, the compliance rate to the high-ability robot, who makes correct decisions 100% of the time, is still less than 100%. As previous research has shown, disuse is a real and common problem in human-automation interaction [10] and often linked to lack of transparency [49]. While we hypothesize that explanations, even offered by a reliable robot, can help improve the trust relationship, compliance rate and team performance, we did not find such an effect in our data from interaction with a high ability robot.

VII. DISCUSSIONS

In this work, we designed an online experiment platform to study trust in HRI. PsychSim was used as the underlying framework to simulate the robot's decision-making process and as the foundation for automatically generated POMDP explanations to establish a proper level of trust. We evaluated these novel explanation algorithms with the testbed where participants teamed up with a simulated robot with either high or low ability, and offered two different types of explanations or no explanations with its decisions. Results indicate that the robot explanations on either confidence-level or observations helped build transparency and trust, and improved decision-making and team performance, particularly so when the robot's ability was low. When the robot's ability was high, the explanations did not make any significant impact on trust, transparency or team performance.

Consistent with previous studies on trust and transparency [5], self-reported trust in the robot's ability was highly correlated with understanding of its decision and decision-making process. However, explanation helped improve understanding of the robot's decision, but only in the low-ability robot. This could be because the high-ability robot always makes correct decisions, so participants never needed to question its decisions, let alone carefully examine its confidence level or observations. Working with a low-ability robot, on the other hand, requires the teammates to pay close attention to the explanations to gauge when to trust or distrust the robot.

This finding on explanations offered by the low-ability robot and subjective trust is seemingly similar to earlier research on hand-crafted explanations [5]. However, in the Dzindolet study, the explanation was provided before the interaction began and was not designed to help participants "diagnose" when to trust the robot's recommendations. Thus, such explanations served more or less as the robot's "excuse" when it was unreliable. The explanations presented here were generated to help participants gauge when and when not to trust the robot. Thus, it is possible that the participants trusted the low-ability robot more when it offered explanations because the robot was more useful, compared to a robot that has the same low ability but did not offer additional information on its decisions.

Interestingly, we did not find any significant differences on the measures we analyzed between confidence-level explanations and observation explanations. Both types of explanations were useful in helping the human teammate decide when to trust the robot. For example, a teammate could potentially learn his/her own heuristics that if the robot's confidence level is below (for example) 75%, then do not follow the robot's decision. Similarly, a teammate could diagnose from the observation explanations that if the camera reports no signs of danger, but the robot's microphone picks up unfriendly conversations, then it is time to be cautious and put protective gear on, regardless of the robot's overall assessment of safety. It is concerning that participants who received confidence-level explanations also felt that they understood the robot's decision-making process, even though such explanations did not reveal any of the robot's inner workings. While confidence-level explanations may help teammates make decisions just as well as with observation explanations, they will not help teammates diagnose or repair the robot (e.g., the participants will not know that it is the camera that caused the robot to make wrong decisions).

Although compliance (e.g., percentage of robot's recommendations followed) is not significantly correlated with mission success, it is significantly correlated with trust in the robot's ability. Additional pairwise correlation tests revealed that compliance is highly correlated with the correct decisions, $r(138) = .957$, $p < .0001$. This is because the robot's errors, although costly, are somewhat rare (16.7%) in the testbed scenario. Future work can vary both the probability and utility of correct decisions.

One of the limitations of the current work is that the understanding of the robot's decision-making process is measured via self-report. In other words, it is unclear whether the participants actually understood such decision-making process, as they claimed. Future work can include measures to test participants' knowledge of the robot (e.g., its capability) or allow it to be inferred more directly and specifically from the subsequent decisions that participants made (e.g., ask participants to choose MOPP gear vs. body armor). Another limitation of the current work is that the measures are aggregated from participants' responses after each of the 3 missions. More fine-grained analysis of data collected from each mission can be conducted to study how trust evolves over time. Additional analysis of individual differences (e.g., complacency potential, uncertainly response) and cognitive load (e.g., NASA's TLX measure) can shed light on how these factors impact the efficacy of explanations on trust, transparency, and team performance. These future analyses can lead to further refinements of our explanation algorithms that can increase the positive impact already exhibited by the current implementation on human-robot trust.

ACKNOWLEDGMENT

This project is funded by the U.S. Army Research Laboratory. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction Studies*, vol. 8, no. 3, pp. 483–500, 2007.

[2] E. Park, Q. Jenkins, and X. Jiang, "Measuring trust of human operators in new generation rescue robots," in *Proceedings of the JFPS International Symposium on Fluid Power*, vol. 2008, no. 7-2. The Japan Fluid Power System Society, 2008, pp. 489–492.

[3] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*. IEEE, 2007, pp. 106–114.

[4] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 719–735, 2003.

[5] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[6] J. Lee and N. Moray, "Trust, self-confidence and supervisory control in a process control simulation," in *Systems, Man, and Cybernetics, 1991.'Decision Aiding for Complex Systems, Conference Proceedings., 1991 IEEE International Conference on*. IEEE, 1991, pp. 291–295.

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

[8] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5, pp. 527–539, 1987.

[9] V. Riley, "Operator reliance on automation: Theory and data," in *Automation and human performance: Theory and applications*, R. Parasuraman and M. Mouloua, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1996, pp. 19–35.

[10] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, 1997.

[11] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[12] W. R. Swartout and J. D. Moore, "Explanation in second generation expert systems," in *Second generation expert systems*. Springer, 1993, pp. 543–585.

[13] L. R. Ye and P. E. Johnson, "The impact of explanation facilities on user acceptance of expert systems advice," *MIS Quarterly*, vol. 19, no. 2, pp. 157–172, 1995.

[14] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[15] A. R. Cassandra, L. P. Kaelbling, and J. A. Kurien, "Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, 1996, pp. 963–972.

[16] S. Koenig and R. Simmons, "Xavier: A robot navigation architecture based on partially observable Markov decision process models," in *Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems*, D. Kortenkamp, R. P. Bonasso, and R. R. Murphy, Eds. MIT Press, 1998, pp. 91–122.

[17] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 271–281, 2003.

[18] S. C. Marsella, D. V. Pynadath, and S. J. Read, "PsychSim: Agent-based modeling of social interactions and influence," in *Proceedings of the International Conference on Cognitive Modeling*, 2004, pp. 243–248.

[19] D. V. Pynadath and S. C. Marsella, "PsychSim: Modeling theory of mind with decision-theoretic agents," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005, pp. 1181–1186.

[20] N. Wang, D. V. Pynadath, K. Unnikrishnan, S. Shankar, and C. Merchant, "Intelligent agents for virtual simulation of human-robot interaction," in *Virtual, Augmented and Mixed Reality*. Springer, 2015, pp. 228–239.

[21] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 2012, pp. 73–80.

[22] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 251–258.

[23] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[24] S. Ososky, D. Schuster, E. Phillips, and F. G. Jentsch, "Building appropriate trust in human-robot teams," in *2013 AAAI Spring Symposium Series*, 2013.

[25] T. Dodson, N. Mattei, and J. Goldsmith, "A natural language argumentation interface for explanation generation in Markov decision processes," in *Algorithmic Decision Theory*, R. I. Brafman, F. S. Roberts, and A. Tsoukiàs, Eds. Springer, 2011, pp. 42–55.

[26] F. Elizalde, L. E. Sucar, M. Luque, J. Diez, and A. Reyes, "Policy explanation in factored Markov decision processes," in *Proceedings of the European Workshop on Probabilistic Graphical Models*, 2008, pp. 97–104.

[27] O. Khan, P. Poupart, and J. Black, "Automatically generated explanations for Markov decision processes," in *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*, L. E. Sucar, E. F. Morales, and J. Hoey, Eds., 2011, pp. 144–163.

[28] V. H. Visschers, R. M. Meertens, W. W. Passchier, and N. N. De Vries, "Probability information in risk communication: a review of the research literature," *Risk Analysis*, vol. 29, no. 2, pp. 267–287, 2009.

[29] L. Hendrickx, C. Vlek, and H. Oppewal, "Relative importance of scenario information and frequency information in the judgment of risk," *Acta Psychologica*, vol. 72, no. 1, pp. 41–63, 1989.

[30] E. A. Waters, N. D. Weinstein, G. A. Colditz, and K. Emmons, "Formats for improving risk communication in medical tradeoff decisions," *Journal of health communication*, vol. 11, no. 2, pp. 167–182, 2006.

[31] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[32] R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? the development of a human–robot interaction trust scale," *International Journal of Social Robotics*, vol. 4, no. 3, pp. 235–248, 2012.

[33] K. E. Schaefer, "The perception and measurement of human-robot trust," Ph.D. dissertation, University of Central Florida Orlando, Florida, 2013.

[34] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 221–228.

[35] P. J. Gmytrasiewicz and E. H. Durfee, "A rigorous, operational formalization of recursive modeling," in *Proceedings of the International Conference on Multi-Agent Systems*, 1995, pp. 125–132.

[36] W. L. Johnson and A. Valente, "Tactical language and culture training systems: Using AI to teach foreign languages and cultures," *Artificial Intelligence Magazine*, vol. 30, no. 2, 2009.

[37] J. M. Kim, J. Randall W. Hill, P. J. Durlach, H. C. Lane, E. Forbell, M. Core, S. Marsella, D. Pynadath, and J. Hart, "BiLAT: A game-based environment for practicing negotiation in a cultural context," *International Journal on Artificial Intelligence in Education: Special Issue on Ill-Defined Domains*, vol. 19, no. 3, pp. 289–308, 2009.

[38] J. Klatt, S. Marsella, and N. Krämer, "Negotiations in the context of AIDS prevention: An agent-based model using theory of mind," in *Proceedings of the International Conference on Intelligent Virtual Agents*, 2011.

[39] R. McAlinden, A. Gordon, H. C. Lane, and D. Pynadath, "UrbanSim: A game-based simulation for counterinsurgency and stability-focused operations," in *Proceedings of the AIED Workshop on Intelligent Educational Games*, 2009.

[40] L. C. Miller, S. Marsella, T. Dey, P. R. Appleby, J. L. Christensen, J. Klatt, and S. J. Read, "Socially optimized learning in virtual environments (SOLVE)," in *Proceedings of the International Conference on Interactive Digital Storytelling*, 2011.

[41] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research*, vol. 13, no. 3, pp. 334–359, 2002.

[42] S. L. McShane. (2014) Propensity to trust scale. http://highered.mheducation.com/sites/0073381225/student_view0/chapter7/self-assessment_7_4.html.

[43] J. M. Ross, *Moderators of trust and reliance across multiple decision aids*. ProQuest, 2008.

[44] D. S. Syrdal, K. Dautenhahn, K. L. Koay, and M. L. Walters, "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," *Adaptive and Emergent Behaviour and Complex Systems*, 2009.

[45] V. Greco and D. Roger, "Coping with uncertainty: The construction and validation of a new measure," *Personality and individual differences*, vol. 31, no. 4, pp. 519–534, 2001.

[46] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[47] R. Taylor, "Situational awareness rating technique(sart): The development of a tool for aircrew systems design," *AGARD, Situational Awareness in Aerospace Operations 17 p(SEE N 90-28972 23-53)*, 1990.

[48] R. C. Mayer and J. H. Davis, "The effect of the performance appraisal system on trust for management: A field quasi-experiment." *Journal of applied psychology*, vol. 84, no. 1, p. 123, 1999.

[49] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-Adapted Interaction*, vol. 18, no. 5, pp. 455–496, 2008.