Building Trust in a Human-Robot Team with Automatically Generated Explanations

Ning Wang, David V. Pynadath
University of Southern California
Los Angeles, CA
nwang@ict.usc.edu, pynadath@usc.edu

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

ABSTRACT

Technological advances offer the promise of robotic systems that work with people to form human-robot teams that are more capable than their individual members. Unfortunately, the increasing capability of such autonomous systems has often failed to increase the capability of the human-robot team. Studies have identified many causes underlying these failures, but one critical aspect of a successful human-machine interaction is trust. When robots are more suited than humans for a certain task, we want the humans to trust the robots to perform that task. When the robots are less suited, we want the humans to appropriately gauge the robots' ability and have people perform the task manually. Failure to do so results in *disuse* of robots in the former case and *misuse* in the latter. Real-world case studies and laboratory experiments show that failures in both cases are common. 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 show that explanations offered by an automated system can help maintain trust with the humans in case the system makes an error, indicating that the robot's communication transparency can be an important factor in earning an appropriate level of trust. To study how robots can communicate their decision-making process to humans, we have designed an agent-based online test-bed that supports virtual simulation of domain-independent human-robot interaction. In the simulation, humans work together with virtual robots as a team. The test-bed allows researchers to conduct online human-subject studies and gain better understanding of how robot communication can improve human-robot team performance by fostering better trust relationships between humans and their robot teammates. In this paper, we describe the details of our design, and illustrate its operation with an example human-robot team reconnaissance task.

ABOUT THE AUTHORS

Ning Wang is a research scientist at the Institute for Creative Technologies of the University of Southern California and the chief scientist at Curious Lab, a consulting service that specializes in assessment of training and simulation technology. Dr. Wang's research is in the design of intelligent virtual characters that motivate students to engage in learning activities. Dr. Wang has extensive experience in the design and assessment of virtual agent based training simulations and educational games. She is one of the pioneers in socially intelligent pedagogical agents – virtual animated characters with human-like behavior that facilitate learning. Her work on the pedagogical agents' socially intelligent feedback lays the groundwork for the design of trust-inducing explanations for the robots in human-robot teams.

David V. Pynadath is the Director for Social Simulation Research at the University of Southern California's Institute for Creative Technologies. He serves as co-chair of the annual Workshop on Plan, Activity, and Intent Recognition. He has published papers on social simulation, multiagent systems, teamwork, plan recognition, and adjustable autonomy. He has developed multiagent systems for applications in social simulation, virtual training environments, automated personal assistants, and UAV coordination. He is the co-creator of PsychSim, a unique multiagent social simulation framework that combines Theory of Mind and decision theory.

Susan G. Hill is a researcher and lead for human-robot interaction at the U.S. Army Research Laboratory, Human Research and Engineering Directorate. She has over 30 years of experience in human factors in a broad range of areas including military systems, human-computer interaction, process control, and industrial ergonomics. Current interests include human interactions with autonomous, intelligent systems in military contexts and how to evaluate such interactions.

Building Trust in a Human-Robot Team with Automatically Generated Explanations

Ning Wang, David V. Pynadath
University of Southern California
Los Angeles, CA
nwang@ict.usc.edu, pynadath@usc.edu

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

INTRODUCTION

Robots have become increasingly intelligent and autonomous and have played important roles in a great variety of task environments, ranging from transportation safety, space exploration, to search and rescue, and many other military operations (Madhavan and Wiegmann, 2007; Li, Rau, and Li, 2010; Bluethmann et al., 2003; Freedy, de Visser, Weltman, & Coeyman, 2007; Burke, Murphy, Coovert, & Riddle, 2004; Kean, 2010; Jones and Schmidlin, 2011; Hinds, Roberts, & Jones, 2004; Parasuraman, Cosenzo, & de Visser, 2009). As robot capabilities grow, the possibility arises that they might provide higher-level functions as full-fledged team members. The assumption that introducing robots into human teams will result in better performance, as compared with when the team or robot operates independently, may not always be justified. Although the addition of robotic systems may lead to improved team capabilities, it may also create challenges that need to be overcome before such hybrid partnerships can work more effectively (Adams et al., 2003).

Ensuring appropriate levels of trust can be a particular challenge to the successful integration of robotic assets in human teams (Freedy et al., 2007). In high-risk or highly uncertain situations, the level of trust in any robotic partner will be critical (Groom & Nass, 2007; Park, Jenkins, & Jiang, 2008). In these contexts, trust can directly affect the willingness of people to accept robot-produced information, follow robots' suggestions, and thus benefit from the advantages inherent in robotic systems (Freedy et al., 2007). The less an individual trusts a robot, the sooner he or she will intervene as it progresses toward task completion (de Visser, Parasuraman, Freedy, Freedy, & Weltman, 2006; Steinfeld et al., 2006). Researchers have also shown that the more operators trust automation, the more they tend to use it. And if operators trust their own abilities more than those of the automated system, they tend to choose manual control (de Vries, Midden, & Bouwhuis, 2003; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & Moray 1991; Lee & Moray 1994; Muir, 1987; Riley 1996). Indeed, trust is a critical element to how humans and robots perform together (Lee & See, 2004). If robots are more suited than humans for a certain task, then we want the humans to trust the robots to perform that task. If the robots are less suited, then we want the humans to appropriately gauge the robots' ability and have people perform the task manually. Failure to do so results in disuse of robots in the former case and misuse in the latter (Parasuraman & Riley, 1997). Real-world case studies and laboratory experiments show that failures in both cases are common (Lee & See, 2004).

As robots gain complexity and autonomy, it is important yet increasingly challenging for humans to understand their decision process. In their book on human-robot interaction in future military operations, Evans and Jentsch (2010) point out that "the key to successful relationship between man and machine is anchored in how well we are able to understand each other and utilize each other's strength while limiting each other's weaknesses". Research has shown 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 (Lee & Moray, 1992). Successful human-robot interaction (HRI) therefore relies on the robot's ability to make its decision-making process transparent to the people it works with. Hand-crafted explanations have shown to be effective in providing such transparency (Dzindolet et al., 2003).

In our work, we pursue a more general approach to explanation that not only builds transparency but can also be reused across domains. We developed an experimental testbed that allows us to quantify the effectiveness of different explanation algorithms in terms of their ability to make a robot's decision-making process transparent to humans. There are several challenges and requirements in the design and implementation of such a testbed. The first challenge is how to model a HRI scenario that facilitates the research of robot communication. A second challenge is the generation of the autonomous behaviors of the robots within that scenario. The robot's decision-making must account for the complex planning, noisy sensors, faulty effectors, etc. that complicate even single-robot execution

and that are often the root of trust failures in HRI. We use a multiagent social simulation framework, PsychSim, as the agent-based platform for our testbed (Marsella, Pynadath, & Read, 2004; Pynadath & Marsella, 2005). Importantly for our purposes, PsychSim includes transparency for explanations that are based in a general decision-theoretic agent framework (Kaelbling, Littman, & Cassandra, 1998). As a result, we have developed a virtual human-robot simulation, where a robot teams up with a human counterpart in reconnaissance missions (Wang, Pynadath, K.V., Shankar, & Merchant, 2015). The robot is modeled as a PsychSim agent, with beliefs and observations of its surroundings, goals (e.g., mission objectives), and actions to achieve those goals. The robot's explanations specifically address three elements that impact trust – Ability, Benevolence and Integrity (Mayer, Davis, & Schoorman, 1995). The design decisions in the implementation of the agent-based online testbed are discussed in detail in our earlier work (Wang et al., 2015). In this paper, we discuss the pilot study on how explanations impact perceived transparency and trust in human robot interactions.

RELATED WORK

While there is abundant research in interpersonal trust and trust in automations, in this paper, we focus on reviewing research on trust in human-robot interactions. Existing theoretical work has proposed a variety of factors that impact trust in HRI. For example, Hancock and colleagues (2011) conducted a meta-analysis of such factors and proposed a triadic model of trust, which categorizes factors of trust as human (e.g., expertise, attitudes towards robots), robot (e.g., reliability, anthropomorphism), and environmental characteristics (e.g., characteristics of the team and task). Similarly, Billings and colleagues (2011) conducted an extensive literature review of human-robot trust and trust between humans and animals to identify overlapping factors between the two and to support the argument that human-animal trust may be an appropriate analog for human-robot trust.

There have also been a growing number of empirical explorations of factors that impact trust in HRI. Freedy et al. (2007) examined how reliability can impact trust using the MITPAS Simulation Environment where participants assumed the role of a controller of an unmanned ground vehicle (UGV). The UGV autonomously targeted and fired, but participants were instructed to take control of the UGV if its behavior would lead to a time delay or a failure. The results suggested that if participants could gauge whether the UGV was very competent or incompetent, they adjusted their behavior accordingly. This adjustment implied that the participants trusted the system to continue to maintain the same level of competence. It was more difficult for users to adjust their behavior when the system showed indeterminate competence. Desai and colleagues (2012) also conducted a series of studies on reliability and trust where participants worked with a robot on a search-and-rescue task. Results show that drops in reliability affected trust, the frequency of autonomy mode switching, as well as participants' self-assessments of performance. In their follow-up work, Desai and colleagues (2013) studied the dynamics of trust during the interaction, to measure the impact not just on the cumulative effect of unreliability, but also on the real-time effect. Their results show that early drops in reliability dramatically lowered real-time trust more than later drops, and the early drops in reliability appeared to promote suboptimal control allocation strategies.

Other studies have shown that people will follow the requests of even an incompetent robot if the negative consequences are somewhat trivial. Salem and colleagues (2015) introduced participants to an error-prone robot who demonstrated its low reliability to them early on. However, participants were still willing to follow the robot's instructions when the tasks did not have high cost (e.g., putting paper in the waste bin). Beyond simply the reliability of the robot, the subjective perceptions that people have of the robot can also influence trust. For example, Ososky and colleagues (2013) used mental model theory to describe how the human team member's understanding of the system contributes to trust in human-robot teaming. They argue that mental models are related to physical and behavioral characteristics of the robot, on the one hand, and affective and behavioral outcomes, such as trust and system use/disuse/misuse, on the other. Another study showed that a robot's personality, particularly how it casts the blame after an error (e.g., to self, teammate or both) can also impact trust in the robot (Kaniarasu & Steinfeld, 2015). Dassonville et al. (1996) conducted a study in which participants used a joystick to control a simulated PUMA arm. Errors were introduced into the simulation, and participants were asked to rate the reliability, performance, and predictability of the joystick's behavior (as well as how difficult it was to make such ratings). The results of the study were consistent with prior work in autonomous systems that suggest that the user's self-confidence is a significant factor which influences use of such systems.

Several survey instruments have been developed to measure trust in HRI (Yagoda, 2011; Schaefer, 2013). For example, Schaefer (2013) conducted a series of studies in human robot team simulations (e.g., joint navigation task). Using data collected from the studies, Schaefer designed a 40-item measure for trust in human-robot interaction. Behavioral measures are frequently used to gauge trust in the robot as well. Prior studies have used a human supervisor's "take-over" and "hand-over" behavior (e.g., takes over a task the robot is currently performing and does it by himself instead) as a measure of the trust or distrust he had in the robot (Xu & Dudek, 2015). We also formulate a rational decision model (described in the next section) that can evaluate trust, similar to Freedy et al. (2007), who 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".

AUTOMATIC GENERATION OF ROBOT EXPLANATION

We have implemented an agent-based online testbed that supports virtual simulation of domain-independent HRI. Our agent framework, PsychSim (Marsella et al., 2004; Pynadath & Marsella, 2005), combines two established agent technologies – decision-theoretic planning (Kaelbling et al., 1998) and recursive modeling (Gmytrasiewicz & Durfee, 1995). Decision-theoretic planning provides an agent with quantitative utility calculations that allow agents to assess tradeoffs between alternative decisions under uncertainty. Recursive modeling gives the agents a *theory of mind* (Whiten, 1991), allowing them to form beliefs about the human users' preferences, factor those preferences into the agent's own decisions, and update its beliefs in response to observations of the user's decisions. The combination of decision theory and theory of mind within a PsychSim agent has proven to be very rich for modeling human decision-making across a wide variety of social and psychological phenomena (Pynadath & Marsella, 2004). This modeling richness has in turn enabled PsychSim agents to operate in a variety of human-agent interaction scenarios (Johnson & Andre, 2009; Kim et al., 2009; Klatt et al., 2011; McAlinden et al., 2009; Miller et al., 2011).

Agent Model

PsychSim agents generate their beliefs and behaviors by solving partially observable Markov decision problems (POMDPs) (Doshi & Perez, 2008; Kaelbling et al., 1998). The POMDP model's quantitative transition probabilities, observation probabilities, and reward functions are a natural fit for our application domain, and they have proven successful in both robot navigation (Cassandra et al., 1996; Koenig & Simmons, 1998) and HRI (Pineau et al., 2003). In our own work, we have used POMDPs to implement agents that acted as 24/7 personal assistants that teamed with researchers to handle a variety of their daily tasks (Chalupsky et al., 2002; Pynadath & Tambe, 2002).

In precise terms, a POMDP is a tuple, $<S, A, T, \Omega, R>$, that we describe in terms of our human-robot team. The *state*, *S*, consists of objective facts about the world, some of which may be hidden from the robot itself. By using a *factored* state representation (Boutilier et al., 2000; Guestrin et al., 2003), the model maintains separate labels and values of each feature of the state, such as the separate locations of the robot and its human teammate, as well as the presence of dangerous people or chemicals in the buildings to be searched. The state also includes feature-value pairs that represent the human teammate's health level, any current commands from the teammate, and the accumulated time cost so far. The robot's available *actions*, *A*, correspond to the possible decisions it can make. Given its search mission, the robot's first decision is where to move to next. We divide the environment into a set of discrete waypoints, so the robot's action set includes potentially moving to any of them. Upon arrival, the robot then makes a decision as to whether to declare a location as safe or unsafe for its human teammate. For example, if the robot believes that armed gunmen are at its current location, then it will 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 in the building. The state of the world changes in response to the actions performed by the robot. We model these dynamics using a *transition probability*, *T,* function that captures the possibly uncertain effects of these actions on the subsequent state. We simplify the robot's navigation task by assuming that a decision to move to a specific waypoint succeeds deterministically. However, we could relax this assumption to decrease the robot's movement ability, as is done in more realistic robot navigation models (Cassandra et al., 1996; Koenig & Simmons, 1998). The robot's recommendation decision affects the health of its teammate, although only stochastically, as there is no guarantee that the teammate will follow the recommendation. Instead, a

recommendation that a building is safe (unsafe) has a high (low) probability of decreasing the teammate's health if there are, in fact, chemicals present.

As already mentioned, the robot and human teammate have only indirect information about the true state of the world. Within the POMDP model, this information comes through a subset of possible *observations*, Ω, that are probabilistically dependent (through the *observation function*, O) on the true values of the corresponding state features. We make some simplifying assumptions, namely that the robot can observe the location of itself and its teammate with no error (e.g., via GPS). However, it cannot detect the presence of armed gunmen or dangerous chemicals with perfect reliability or omniscience. Instead, it receives a local reading about their presence (or absence) at its current location. For example, if dangerous chemicals are present, then the robot's chemical sensor will detect them with a high probability. However, there is also a lower, but nonzero, probability that the sensor will *not* detect them. In addition to such a false negative, we can also model a potential false positive reading, where there is a low, but nonzero, probability that it will detect chemicals even if there are none present. By controlling the observations that the robot receives, we can manipulate its *ability* in our testbed. Partial observability gives the robot only a subjective view of the world, where it forms beliefs about what it *thinks* is the state of the world, computed via standard POMDP *state estimation* algorithms. For example, the robot's beliefs include its subjective view on the presence of threats, in the form of a likelihood (e.g., a 67% chance that there are toxic chemicals in the farm supply store). Again, the robot derives these beliefs from its local sensor readings, so they may diverge from the true state of the world. By decreasing the accuracy of the robot's observation function, O, we can decrease the accuracy of its beliefs, whether receiving correct or incorrect observations. 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 the human-robot team's mission objectives as a *reward*, R, that maps the state of the world into a real-valued evaluation of benefit for the agent. The highest reward is earned in states where all buildings have been explored by the human teammate. This reward component incentivizes the robot to pursue the overall mission objective. There is also an increasingly positive reward associated with level of the human teammate's health. This reward component punishes the robot if it fails to warn its teammate of dangerous buildings. Finally, there is a negative reward that increases with the time cost of the current state. This motivates 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 current beliefs about the state of the world (Kaelbling et al., 1998). Rather than perform an offline computation of a complete optimal policy over all possible beliefs, we instead take an online approach so that the robot makes optimal decisions with respect to only its current beliefs (Ross et al., 2008). The robot uses a bounded lookahead procedure that seeks to maximize expected reward by simulating the dynamics of the world from its current belief state. In particular, the robot will consider declaring a building dangerous or safe (i.e., recommending that its teammate put protective gear on or not). It will 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 impact on the teammate's health and the time to search the building. It will finally combine these outcome likelihoods with its reward function and choose the option that has the highest reward.

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 agents or robots. To begin with, PsychSim agents provide support for transparent reasoning that is a requirement for our testbed. PsychSim makes the agent's reasoning process available to the developer, in the form of a branching tree representing its expected value calculation. In this work, we apply this capability to manipulate the explanations that the robot gives to its human teammate. 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 the contents of its model into human-readable sentences:

- *S*: The robot can communicate its beliefs directly to the user, e.g., "There is a 67% probability that dangerous chemicals are present in this building."
- *A*: The robot can instead make a decision as to 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."
- *P*: The robot can also reveal the relative likelihood of possible outcomes, e.g., "There is a 67% probability that you will be injured if you enter the doctor's 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 informants house without putting on protective gear. The protective gear will slow you down a little."

STUDY WITH HUMAN ROBOT INTERACTION ONLINE TESTBED

We developed an online HRI simulation testbed to study the design of automatically generated explanations to influence trust. The testbed can be accessed from a web browser. The current testbed implements a scenario in which a human teammate works with a robot in reconnaissance missions to gather intelligence in a foreign town. The mission involves the human teammate 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 gunman, and a microphone that can listen to discussions in foreign language. It determines whether danger may be present if its human teammate enters the building.

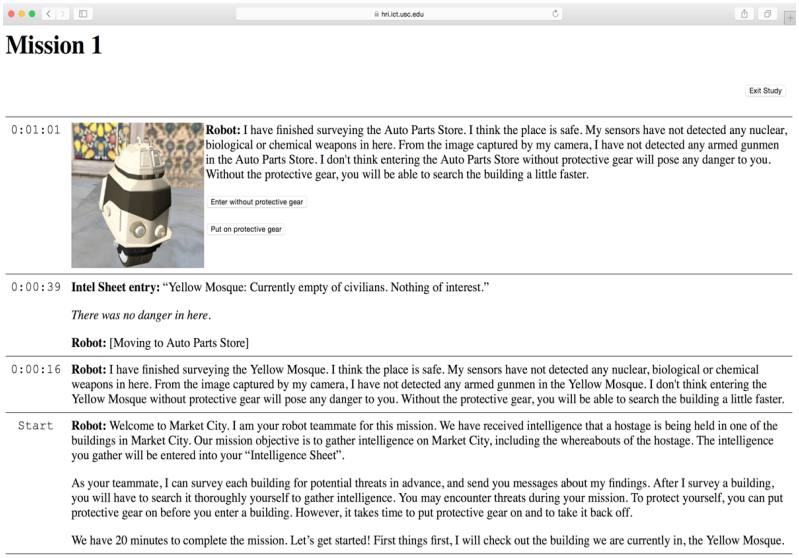

Figure 1. Online HRI Testbed.

We used the online testbed to conduct a pilot study on how the robot's explanation impact trust. We designed four versions of the simulated robot, varied along two dimensions – ability (high vs. low) and explanation (with or without). The study is a between-subject design. Each participant interacted with one of the four simulated robots. The robot with high ability makes the correct decision 100% of the time. The one with low ability has a faulty camera and only makes false-negative mistakes, e.g., not detecting armed gunmen in the simulation. The other simulated sensors (e.g., NBC weapon detector and microphone) and the robot's decision-making capability remain intact. When the robot communicates without explanation, it tells its teammate (e.g., the participant) only its

decisions (e.g., "I have finished surveying the doctor's office. I think the place is safe."). In this study, only the *A* component from PsychSim is included in the decision. When the robot communicates with explanations, it will communicate the decision and the explanations. In this study, the explanations focus on the robot's ability, particularly its sensing capability. Thus the and *R* components (as well as *A*) from PsychSim are included in the explanation. The explanation is designed to help the robot's teammate understand which sensors are working correctly and which ones are not. Additionally, the explanations communicate the robot's capability to understand the human teammate's goals. Together, the explanation is designed to influence *ability*, *benevolence* and potentially *integrity* dimensions of trust (Mayer et al., 1995). One such communication with both decision and explanation from our scenario would be: *"I have finished surveying the Cafe. I think the place is dangerous. 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. I think it will be dangerous for you to enter the Cafe without protective gear. The protective gear will slow you down a little."*

Participants

We recruited 120 participants from Amazon Mechanical Turk (AMT). However, 1 additional participant gained access to the study website. As a result, data from 121 participants are included. The participants have previously completed 500 or more jobs on AMT and have a completion rate of 95% or higher. All participants are located in the United States. The participants average 33 years old. 42% of the participants are female and 58% participants are male. None of the participants are active service members. Only 3 participants have answered that they have worked with an automated squad member before. Only 1 participant had reconnaissance or search and rescue training.

Procedure

Each participant first read an information sheet about the study and then filled out the background survey. Next, participants worked with a simulated robot on three reconnaissance missions. After each mission, participants filled out a post-mission survey. Each participant worked with the robot with the same ability and communication (e.g., low ability and communicates with explanations) throughout the three missions. Participants were randomly assigned to team up with one of the four robot conditions. 31 participants interacted with a robot with high ability and no explanation. 29 participants interacted with a robot with high ability and explanation. 30 participants interacted with a robot with low ability and no explanation and 31 interacted with a robot with low ability and explanation. The study was designed to be completed in 90 minutes.

Measures

The *Background Survey* includes measures of the demographic information, education, video game experience, military background, predisposition to trust (McKnight, Choudhury, & Kacmar, 2002), propensity to trust (McShane, 2015), complacency potential (Ross, 2008), negative attitude towards robots (Syrdal, Dautenhahn, Koay, & Walters, 2009) and uncertainty response scale (Greco & Roger, 2001). In the *Post-Mission Survey*, we have designed items to measure participants' understanding of the robot's decisions and decision-making process. We modified items on interpersonal trust to measure trust in the robot's ability, benevolence and integrity (Mayer & Davis, 1999). We also included the NASA Cognitive Load Index (Hart & Staveland, 1988), Situation Awareness Rating Scale (Taylor, 1990), trust in oneself and teammate (Ross, 2008), and trust in robots (Schaefer, 2013). We have also collected *interaction logs* from the online testbed. In this paper, we present a preliminary analysis of the date collected. The analysis includes measures on mission success derived from the *interaction log*, and trust in the robot's ability (Mayer & Davis, 1999) and the scale we designed on the understanding of the robot's decision-making process presented in the *Post Mission Survey*. A more comprehensive analysis that includes the full specturm of the measures will be part of our future work.

RESULTS

As a measure of trust in the robot's ability, we averaged the participants' self-report on trust in the robot's ability after each mission. We adopted the same method to measure participants' understanding of the robot's decision and decision-making process. Overall, the experiment manipulation of the robot's ability was successful. Each participant took part in 3 missions. A one-way ANOVA test shows that participants who interacted with a high-

ability robot completed their missions more often, compared to participants who interacted with a low-ability robot ($M_{high\_ability} = 2.6$, $M_{low\_ability} = 1.7$, $p < .001$). Not surprisingly, participants who interacted with a high-ability robot also had a higher level of trust in the robot's ability ($M_{high\_ability} = 6.8$, $M_{low\_ability} = 4.6$, $p < .001$, on a 1-7 Likert scale). Surprisingly though, participants who interacted with a high-ability robot reported that they understood the robot's decisions and decision-making process more than participants who interacted with a low-ability robot ($M_{high\_ability} = 6.4$, $M_{low\_ability} = 4.4$, $p < .001$, on a 1-7 Likert scale). We did not find a significant main effect of the robot's explanation on mission success and trust in the robot's ability. However, participants who interacted with a robot that offered explanations reported that they understood the robot's decisions and decision-making process more than participants who interacted with a robot that did not offer explanations with its decisions ($M_{with\_explanation} = 5.8$, $M_{without\_explanation} = 5.2$, $p = .04$). We did not find any significant interaction between the robot's ability and explanation on mission success, trust in the robot's ability, and understanding of the robot's decisions.

Because the investigation in this paper focuses primarily on the impact of the robot's explanation when the robot's ability varies, we broke down the comparison into two groups: high- and low-ability robot. Within the group of participants who interacted with a low-ability robot, we did not find any significant difference on mission success and trust in the robot's ability between participants who interacted with a robot that gave explanations and participants who interacted with a robot that did not give explanations. However, participants who interacted with a robot that provided explanations along with its decisions reported that they understood the robot's decisions and decision-making process more ($M_{with\_explanation} = 5.0$, $M_{without\_explanation} = 4.2$, $p = .03$). Within the group of participants who interacted with a high-ability robot, we did not find significant impact of the robot's explanations on mission success and understanding of the robot's decisions. We did find a marginally significant difference in the impact of the robot's explanations on trust in the robot's ability. Particularly, participants who received the robot's explanations reported that they trust the robot's ability more, compared to participants who did not receive explanations with the robot's decisions ($M_{with\_explanation} = 6.9$, $M_{without\_explanation} = 6.6$, $p = .0512$).

DISCUSSION

In this paper, we discussed the design of an online experiment platform to study trust in human-robot interactions. PsychSim was used as the underlying framework to simulate the robot's decision-making process and as the foundation for automatically generated robot explanations to establish a proper level of trust. We conducted a pilot study with the testbed where participants teamed up with a simulated robot with either high or low ability, and offered explanations or no explanations with its decisions. Results from our preliminary analysis indicate that the experiment testbed design was valid in that the participants were able to complete the mission with help from the robot, and more successfully doing so when working with a high-ability robot. The robot explanation did not help participants complete the mission more successfully or trust the robot's ability more. However, the explanations did successfully establish perceived transparency: participants felt that they understood the robot's decision and decision making better when explanations were provided along with decisions, particularly so when the robot's ability is low. Interestingly, when the robot's ability was high, explanations made the participants feel that they could trust the robot's ability, more so than when the robot's decisions were delivered without explanations.

The results help us gain interesting insight into how to design explanations to establish trust. Previous studies have shown that transparency can lead to trust (Dzindolet et al., 2003). That result was the inspiration and rationale for us to design robot explanations to build transparency. However, transparency alone may not be sufficient to establish trust. In our study, explanations improved perceived transparency and made the participants feel that they understood the robot's decisions and decision-making process. However, the understanding did not help the participants make better decisions during the mission. For example, the explanations were designed to help participants understand that the robot's camera is the only sensor that is not functioning properly and will not be able to detect armed gunmen as a result. Understanding that the robot's limited sensing capability may lead to false-negative detection of danger still left participants at a loss of what to do. While the safest solution would be to always equip protective gear, participants were only given limited time to complete the mission and equipping protective gear costs time. In the end, the explanation did not help participants decide – should I put protective gear on when the robot decides the building I am about to enter is safe? This is reflected in the result that overall the explanations did not impact the participants' trust in the robot's ability because with or without explanation, the low-ability robot is still not helpful, and the robot's ability alone influenced their trust in the robot. Particularly, items in the measure of the trust in robot's ability (Mayer & Davis, 1999), such as "the robot is capable of its job" and "the

robot has specialized capability that can increase our performance", are tightly connected to outcome and performance. It is important to note that overall, the participants' trust in a low-ability robot was much lower than the participants' trust in a high-ability robot. And offering explanations did not (and *should* not) make the participants perceive the low-ability robot as more trustworthy.

One of the limitations of the current work is that the understanding of the robot's decisions is measured via self-report. In other words, it is unclear whether the participants actually understood the decisions, 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 equipping MOPP gear vs. body armor.

The results also suggest that explanations can help establish trust in the robot's ability, even when the robot is fully capable of making correct decisions 100% of the time. This indicates that explanations can potentially alleviate the problem of disusing or under-utilizing the robot, usually as a result of the lack of trust. In our study, the explanation did not help participants achieve more mission success, when working with the high-ability robot. However, it could be due to the ceiling effect, e.g., there were only 3 missions and the participants succeeded on average in more than 2 of them (average 2.6). Future studies where this "ceiling" is raised can help shed light on the impact of explanations on the disuse of the robot.

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. We have also developed a variation of the HRI online testbed in an immersive 3D simulation. The 3D HRI simulation testbed aims to enhance the interaction fidelity to allow the interaction to be closer to physical human-robot interaction, and the robot's explanations to be more similar to that of a physical robot and less so to a message from an automated system. The data collected from the current study will inform the refinement and future study with the 3D HRI simulation testbed.

ACKNOWLEDGEMENTS

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

Adams, B. D., Bruyn, L. E., Houde, S., & Angelopoulos, P. (2003). Trust in automated systems literature review (DRDC Toronto No. CR-2003-096). Toronto, Canada: Defence Research and Development Canada.

Billings, D. R., Schaefer, K. E., Chen, J. Y., Kocsis, V., Barrera, M., Cook, J., Ferrer, M., & Hancock, P. A. (2012). Human-animal trust as an analog for human-robot trust: A review of current evidence (No. ARL-TR-5949). UNIVERSITY OF CENTRAL FLORIDA ORLANDO.

Bluethmann, W., Ambrose, R., Diftler, M., Askew, S., Huber, E., Goza, M., Rehnmark, F., Lovchik, C., Magruder, D. (2003). Robonaut: A Robot Designed to Work With Humans in Space. Autonomous Robots, 14, 34–39.

Boutilier, C., Dearden, R., & Goldszmidt, M. (2000). Stochastic dynamic programming with factored representations. Artificial Intelligence, 121(1), 49-107.

Burke, J., Murphy, R., Coovert, M., Riddle, D. (2004). Moonlight in Miami: An Ethnographic Study of Human-Robot Interaction in USAR. Human-Computer Interaction, 19 (1/2), 85–116.

Cassandra, A. R., Kaelbling, L. P., & Kurien, J. (1996, November). Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In Intelligent Robots and Systems' 96, IROS 96, Proceedings of the 1996 IEEE/RSJ International Conference on (Vol. 2, pp. 963-972). IEEE.

Chalupsky, H., Gil, Y., Knoblock, C. A., Lerman, K., Oh, J., Pynadath, D. V., Ross, T. A., & Tambe, M. (2002). Electric Elves: Agent technology for supporting human organizations. AI magazine, 23(2), 11.

Dassonville, I., Jolly, D., & Desodt, A. M. (1996). Trust between man and machine in a teleoperation system. Reliability Engineering & System Safety, 53(3), 319-325.

de Visser, E. J., Parasuraman, R., Freedy, A., Freedy, E., & Weltman, G. (2006). A comprehensive methodology for assessing human-robot team performance for use in training and simulation. In Proceedings of the 50th

Annual Meeting of the Human Factors and Ergonomics Society (pp. 2639–2643). Santa Monica, CA: Human Factors and Ergonomics Society.

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human-Computer Studies, 58(6), 719-735.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012, March). Effects of changing reliability on trust of robot systems. In Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on (pp. 73-80). IEEE.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). Impact of robot failures and feedback on real-time trust. In Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction (pp. 251-258). IEEE Press.

Doshi, P., & Perez, D. (2008, July). Generalized Point Based Value Iteration for Interactive POMDPs. In AAAI (pp. 63-68).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697-718.

Evans, A. W., & Jentsch, F. (2010). The future of HRI: Alternate research trajectories and their influence on the future of unmanned systems. Human-Robot Interactions in Future Military Operations (Eds. Mike Barnes and Florian Jentsch). Ashgate Publishing, Surrey, UK.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human-robot collaboration. In Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on (pp. 106-114). IEEE.

Gmytrasiewicz, P. J., & Durfee, E. H. (1995, June). A Rigorous, Operational Formalization of Recursive Modeling. In ICMAS (pp. 125-132).

Greco, V., & Roger, D. (2001). Coping with uncertainty: The construction and validation of a new measure. Personality and Individual Differences, 31, 519–534.

Groom, V., & Nass, C. (2007). Can robots be teammates? Benchmarks in human-robot teams. Interaction Studies, 8, 483–500.

Guestrin, C., Koller, D., Parr, R., & Venkataraman, S. (2003). Efficient solution algorithms for factored MDPs. Journal of Artificial Intelligence Research, 399-468.

Hancock, P. A., Billings, D. R., Schaefer, K., Chen, J. Y. C., de Visser, E. J., Parasuraman, R. (2011) A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors, 53 (5), 517–527.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology, 52, 139-183.

Johnson, W. L., & Valente, A. (2009). Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. AI Magazine, 30(2), 72.

Jones, K. S., Schmidlin, E. A. (2011) Human-Robot Interaction: Toward Usable Personal Service Robots. Reviews of Human Factors and Ergonomics, 7, 100–148.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial intelligence, 101(1), 99-134.

Kaniarasu, P., & Steinfeld, A. M. (2014, August). Effects of blame on trust in human robot interaction. In Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on (pp. 850-855). IEEE.

Kean, S. (2010) Making Smarter, Savvier Robot. Science, 329, 508–509.

Kim, J. M., Hill Jr, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., Marsella, S. C., Pynadath, D. V., & Hart, J. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. International Journal of Artificial Intelligence in Education, 19(3), 289-308.

Klatt, J., Marsella, S., & Krämer, N. C. (2011, January). Negotiations in the context of AIDS prevention: an agent-based model using theory of mind. In Intelligent Virtual Agents (pp. 209-215). Springer Berlin Heidelberg.

Koenig, S., & Simmons, R. (1998). Xavier: A robot navigation architecture based on partially observable markov decision process models. Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems, 91-122.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International journal of human-computer studies, 40(1), 153-184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50-80.

Lee, J., & Moray, N. (1991, October). 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 (pp. 291-295). IEEE.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270.

Li, D., Rau, P. L. P., Li, Y. (2010) A Cross-Cultural Study: Effect of Robot Appearance and Task. International Journal of Social Robotics, 2, 175–186.

Madhavan, P., Wiegmann, D. A. (2007). Similarities and Differences Between Human-Human and Human-Automation Trust: An Integrative Review. Theoretical Issues in Ergonomics Science, 8 (4), 277–301.

Marsella, S. C., Pynadath, D. V., & Read, S. J. (2004). PsychSim: Agent-based modeling of social interactions and influence. In Proceedings of the international conference on cognitive modeling (Vol. 36, pp. 243-248).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

McAlinden, R., Gordon, A., Lane, H. C., & Pynadath, D. (2009). UrbanSim: A game-based simulation for counterinsurgency and stability-focused operations. In Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK (pp. 41-50).

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information systems research, 13(3), 334-359.

McShane, S. L., (2014). Propensity to Trust Scale. Retrieved from http://highered.mheducation.com/sites/0073381225/student_view0/chapter7/self-assessment_7_4.html

Miller, L. C., Marsella, S., Dey, T., Appleby, P. R., Christensen, J. L., Klatt, J., & Read, S. J. (2011). Socially optimized learning in virtual environments (SOLVE). In Interactive Storytelling (pp. 182-192). Springer Berlin Heidelberg.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies, 27(5), 527-539.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013, March). Building appropriate trust in human-robot teams. In 2013 AAAI Spring Symposium Series.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52, 381–410.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230-253.

Park, E., Jenkins, Q., & Jiang, X. (2008 September). Measuring trust of human operators in new generation rescue robots. Paper presented at the 7th JFPS International Symposium on Fluid Power, Toyama, Japan.

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. Robotics and Autonomous Systems, 42(3), 271-281.

Pynadath, D. V., & Marsella, S. C. (2004, July). Fitting and compilation of multiagent models through piecewise linear functions. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3 (pp. 1197-1204). IEEE Computer Society.

Pynadath, D. V., & Marsella, S. C. (2005, July). PsychSim: Modeling theory of mind with decision-theoretic agents. In IJCAI (Vol. 5, pp. 1181-1186).

Pynadath, D. V., & Tambe, M. (2002). Electric Elves: Adjustable Autonomy in Real-world Multiagent Environments. In Socially Intelligent Agents: Creating Relationships with Computers and Robots. 12, 101-108, Kluwer: Alphen aan den Rijn, Netherlands

Riley, V. (1996). Operator reliance on automation: Theory and data.

Ross, J. M. (2008). Moderators of trust and reliance across multiple decision aids (Doctoral dissertation). University of Central Florida, Orlando.

Ross, S., Pineau, J., Paquet, S., & Chaib-Draa, B. (2008). Online planning algorithms for POMDPs. Journal of Artificial Intelligence Research, 663-704.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, March). 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 (pp. 141-148). ACM.

Schaefer, K. E. (2013). The perception and measurement of human-robot trust (Doctoral dissertation, University of Central Florida Orlando, Florida).

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. In Proceedings of 2006 ACM Conference on Human-Robot Interaction, (pp. 33–40). Salt Lake City, UT: ACM.

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). 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.

Taylor, R. M. (1990). 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).

Wang, N., Pynadath, D. V., K.V., U., Shankar, S., & Merchant, C. (20015). Intelligent Agents for Virtual Simulation of Human-Robot Interaction. To appear in Proceedings of the 17th International Conference on Human-Computer Interaction.

Whiten, A. (Ed.). (1991). Natural theories of mind: Evolution, development and simulation of everyday mindreading. Oxford: Basil Blackwell.

Xu, A., & Dudek, G. (2015, March). 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 (pp. 221-228). ACM.

Yagoda, R. E. (2011). WHAT! You want me to trust a ROBOT? The development of a human robot interaction (HRI) trust scale (Master's Thesis). NC State University.