Is It My looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams

Ning Wang[1], David V. Pynadath[1], Ericka Rovira[2], Michael J. Barnes[3], Susan G. Hill[3]

[1]Institute for Creative Technologies, University of Southern California,
[2]U.S. Military Academy, West Point, [3]U.S. Army Research Laboratory

Abstract. Trust is critical to the success of human-robot interaction. Research has shown that people will more accurately trust a robot if they have an accurate understanding of its decision-making process. The Partially Observable Markov Decision Process (POMDP) is one such decision-making process, but its quantitative reasoning is typically opaque to people. This lack of transparency is exacerbated when a robot can learn, making its decision making better, but also less predictable. Recent research has shown promise in calibrating human-robot trust by automatically generating explanations of POMDP-based decisions. In this work, we explore factors that can potentially interact with such explanations in influencing human decision-making in human-robot teams. We focus on explanations with quantitative expressions of uncertainty and experiment with common design factors of a robot: its embodiment and its communication strategy in case of an error. Results help us identify valuable properties and dynamics of the human-robot trust relationship.

1 Introduction

Trust is critical to the success of human-robot interaction (HRI) [1]. To maximize the performance of human-robot teams, people should trust their robot teammates to perform a given task when robots are more suited than humans for the task. If the robots are less suited, then people should perform the task themselves. Failure to do so results in *disuse* of robots in the former case and *misuse* in the latter [2]. Real-world case studies and laboratory experiments show that failures of both types are common [3].

Research has shown that people will more accurately trust an autonomous system if they have a more accurate understanding of its decision-making process [4]. The Partially Observable Markov Decision Process (POMDP) is one such decision-making process, providing optimized decision making that is commonly used by robots, agents, and other autonomous systems [5]. Unfortunately, the quantitative nature of POMDP algorithms and their results makes them hard for people to understand. Furthermore, while a robot could learn to improve its POMDP model, such changes in its decision-making only exacerbate the lack of transparency. Fortunately, recent research has shown promise in calibrating human-agent trust by automatically generating explanations of POMDP-based decisions [6].

In this work, we seek a deeper understanding of the factors leading to the effectiveness (or lack thereof) of such automatically generated explanations. We specifically focus on explanations that provide quantitative information on uncertainty and two factors related to common robot design decisions: its embodiment and its communication strategy in case of errors. We seek to understand how a robot's coping strategies after making an error may interact with its transparency communications in calibrating a human teammate's trust in it. We implement a specific trust-repair strategy inspired by prior work in organizational trust: an acknowledgement of a mistake, paired with a promise to improve [7]. We thus can study differences in the effect of such an error acknowledgment and promise to learn when preceded by different types of explanations of the robot's decision-making.

In addition, people have been observed to react differently to robot teammates based on their appearance [8]. There are clear behavioral differences for many people when interacting with more human- or animal-like robots, in contrast to their interactions with more "mechanical" robots. In fact, trust in human-animal interaction shares some characteristics with trust in human-robot interaction, in that both seek to augment human skills and abilities in order to better accomplish a particular task [9]. It has been suggested that human-animal interactions may represent a suitable metaphor for human-robot interactions (for review, see [10]). Of course, the roles that each entity fills depend on its capabilities, skills, and affordances [11, 12]. Thus we consider how the robot's embodiment will affect the interpretation and effectiveness of its explanations. In particular, we draw inspiration from studies showing that dog-like robots are treated differently from those with a more traditionally robotic appearance [11, 12]. We can therefore quantify the potentially different effects of POMDP-based explanations when coming from robots with different embodiments.

To quantify the impact of these variables, we expand our measures to consider self-reported trust as well as behavioral measures of human decision-making, such as compliance with the robot's recommendations, correct decisions by the human teammate, and correct diagnosis of the robot's failures by the human teammate. By looking at where these behavioral measures deviate from self-reported measures, we can better drill down into the mental states of the human teammates, into the antecedents of trust.

2 Related Work

Existing studies have shown that a human's ability to understand its agent teammate has a clear impact on trust [4]. Hand-crafted explanations have shown to contribute to that understanding in a way that provides transparency and improves trust [13]. Automated, domain-independent methods for generating explanations have a long history within the context of rule- and logic-based systems, like expert systems [14]. There has been more recent work on generating explanations based on Markov Decision Problems (MDPs) [15]. Our previous work automatically generated explanations from Partially Observable MDPs [6], which provide a more realistic model for HRI domains, due to the inherent uncertainty in the robots' operating environment. The existing evidence is encouraging as to the potential success of applying a general-purpose explanation on top of an agent's decision-making process.

To identify the most effective content for such AI-based explanations, we look to 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 indicates that "people prefer numerical information for its accuracy but use a verbal statement to express a probability to others." [16]. 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 [17]. 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 [18]. In translating our robot's POMDP-based reasoning into a human-understandable format, our explanation algorithms use natural-language templates inspired by these various findings in the literature.

Previous studies of automatically generated explanations in HRI used a fixed robot, with a traditionally mechanical appearance. However, people react differently to robot teammates based on their appearance. Prior studies have shown that some people show a marked preference for more mechanical-looking robots, while others are more comfortable interacting with humanoid robots [8]. Such observations have prompted other technological attempts to emulate the physical, behavioral, and cognitive aspects of biological entities within robots. Studies have found that humans tend to describe their relationships with robotic animals as similar to those with biological animals [11, 12]. These studies found that people will often attribute some (but not all) dog-like qualities to robots who look like dogs. In fact, in many domains, human-animal trust can be viewed as a better model for HRI than human-human trust [10].

Given the uncertain nature of the robot's decisions, they will inevitably turn out to be wrong from time to time. Reinforcement learning has enabled many robots to improve from their mistakes (e.g., [19, 20]). While such learning is likely to complicate the robot's effort to reason transparently with human teammates, it does provide an opportunity to repair trust that has been damaged by robot errors. Our investigation into the interaction between explanations and trust repair is inspired by work on the latter within organizations [21, 7]. Prior research has found that timely trust-repair actions are critical to effectively maintaining trust within HRI [22].

In this paper, we discuss the impact of a robot's embodiment, its explanation, and its promise to learn from mistakes on trust and team performance. Based on results from previous studies of robot explanations and trust [23], we hypothesize that:

H1: Compared to a robot who offers no explanations of its decisions, a robot who offers explanations can help its teammate better calibrate trust and produce better team performance.

Additionally, we hypothesize that a robot that looks like an animal, such as a dog, will help human teammates establish a trust relationship with it similar to the one they would have with a real dog. We therefore hypothesize that:

H2: A robot's embodiment will impact trust in the robot and team performance. Specifically, a robot whose appearance shares that of an animal will foster a stronger trust relationship than one with a more mechanical appearance.

Finally, a robot can acknowledge its mistakes and promise to learn from them, so as to indicate that it is aware of its limitations and knows how to improve. Such indications can potentially improve its teammate's trust in its ability. We hypothesize that:

H3: A robot that acknowledges each mistake it makes and promises to learn from them will improve its trust relationship with its human teammates.

3 HRI Testbed

We evaluate our hypotheses in the context of an online HRI testbed [24]. For the current study, we used the testbed to implement a scenario in which a human teammate works with a different robot across eight reconnaissance missions (Figure 1). Each mission requires the human teammate to search 15 buildings in a different town. The virtual robot serves as a scout, scans the buildings for potential danger, and relays its findings. The robot has an NBC (nuclear, biological, and chemical) weapon sensor, a camera that can detect armed gunmen, and a microphone that can identify suspicious conversations.

The human must choose between entering a building with or without protective gear. If there is danger in the building, the human will be injured if not wearing the protective gear, and the team will incur a 3-minute time penalty. However, it takes time to put on and take off protective gear (20 seconds each). The human teammate must enter all 15 buildings within 10 minutes; otherwise, the mission is a failure. So the human is incentivized to consider the robot's findings before deciding how to enter the building.

We model this task as a POMDP, which is a tuple, $\langle S, A, P, \ \Omega, O, R \rangle$ [5]. The state, S, consists of objective facts about the world, such as the presence of dangerous chemicals in the buildings. The robot's available actions, A, correspond to the possible decisions it can make. Upon arrival at a new building, the robot makes a decision as to whether to declare it safe or unsafe for its human teammate. We model the dynamics of the world using a transition probability function, P, that captures the uncertain effects of the robot's actions. A recommendation that a building is safe (unsafe) has a high (low) probability of decreasing the teammate's health if there is, in fact, danger present.

The robot has only indirect information about the true state of the world, 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. For example, if dangerous chemicals are present at its current location, 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.

The robot's reward, R, is highest when all buildings have been explored by the human teammate. This reward component incentivizes the robot to pursue the overall mission objective. There is also a positive reward associated with 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 quickly.

An agent can generate behavior based on its POMDP model by determining the optimal action based on its current beliefs, b, about the state of the world [5]. In particular, our robot will consider declaring a building dangerous or safe by combining its beliefs about the likelihood of possible threats in the building with each possible declaration

to compute the likelihood of the outcome (i.e., impact on teammate's health and time needed to search the building). It will finally combine these outcome likelihoods with its reward function and choose the option that has the highest reward.

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 procedures for searching buildings. The participant needs to decide only whether to trust the robot's findings (safe/dangerous) and press a button to enter/exit the room. In the current study, we fixed the observations the robot receives to be accurate 80% of the time. As a result, the robot makes incorrect assessment of the danger level for 3 out of 15 buildings in each town. Research on automation reliability on trust and human-automation team has indicated that a reliability of 80% ([25]) or above 70% ([26]) to be a suitable setting for similar studies.

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

4 Evaluation

The domain of the testbed scenario is relevant to the military, so we recruited 61 participants from a higher-education military school in the United States. Participants were awarded extra course credit for their participation.

Design: We varied the robot's embodiment (robot vs. robot dog), explanation (no explanation vs. confidence explanation) and acknowledgement to learn from mistakes (no acknowledgement vs. acknowledgement). The aforementioned testbed was used in the study. Because individual differences often impact trust in automation [27], a $2 \times 2 \times 2$ within-subject design is used in the study. Each participant completed 8 missions. In each mission, a different variation of the robot worked with the participant. A total of 8 variations of robot were used (hence 8 missions). While the task and environment of mission 1 through 8 were fixed, the order of the robot variations was counter-balanced. At the beginning of each mission, participants were told that they were working with a new robot for the first time (e.g., not the same robot from previous missions).

Embodiment: Two robot embodiments were used in the study, illustrated in Figure 2. One robot was designed to look like a dog, with ears, nose, and highlighted eyes, suggesting possibly embedded sound, NBC, and vision sensors. The second robot was designed to have the appearance of a typical robot-looking robot on wheels.

Explanation: Existing algorithms explain an agent's decision-making by exposing different components of its POMDP model [6]. In this study, the explanation variable

has two levels: no explanation and a confidence-level explanation. At both levels, the robot informs its teammate of its decision (e.g., "I have finished surveying the doctor's office. I think the place is safe."). Under the confidence-level explanations, the robot augments this decision with additional information that should help its teammate better understand its ability (e.g., decision-making), one of the key dimensions of trust [28]. The confidence-level explanations augment the decision message with additional information about the robot's uncertainty in its decision. 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 robot's one faulty sensor will lead to occasionally conflicting observations, it will on those occasions have lower confidence in its erroneous decisions.

Fig. 2. The two embodiment of the robots used in the study: a robot (left) and a robot dog (right).

Acknowledgment: The acknowledgement variable has two levels: no acknowledgement and an acknowledgement that a mistake has been made along with a promise to learn from the mistake. This acknowledgement is given every time the robot makes an assessment that turned out to be incorrect. The team searches 15 buildings during each reconnaissance mission. In each mission, the robot makes an incorrect assessment of three buildings. An example of the robot's acknowledgement is "It seems that my assessment of the informant's house was incorrect. I will update my algorithms when we return to base after the mission."

Procedure Participants first read an information sheet and filled out the online background survey. Next, participants worked with a simulated robot on 8 reconnaissance missions. In each mission, a variation of the simulated robot (with a different combination of embodiment, explanation, and acknowledgment to learn from its mistakes) was presented. The order in which the robots were presented was counter-balanced across participants. After each mission, participants filled out an online post-mission survey. The study was designed to be completed in 2 sessions, 120 minutes total.

Measure The Background Survey included measures of demographic information, education, video game experience, military background, predisposition to trust [29], propensity to trust [30], complacency potential [31], negative attitude towards robots [32], and the uncertainty response scale [33]. Because the impact of individual differences on trust is not the focus of this paper, such analyses and results are not included.

In the Post-Mission Survey, we 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 [28]. We also included

the NASA Cognitive Load Index [34], Situation Awareness Rating Scale [35], and trust in oneself and teammate [31]. We have also collected interaction logs from the testbed.

The dependent measures discussed in this paper are listed below. Trust can be measured via both self-report [28] 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, because 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 was measured by modifying an existing scale [36]. Each factor of trust was 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 *benevolence* and *integrity*. So we focus on only the *ability* component of trust in this paper. The value ranges from 1 to 7.

Transparency: This is measured using items (along a 1–7 Likert scale) on the understanding of the robot's decision-making process, designed by the researchers. A sample item from this measure is "I understand the robot's decision-making process".

Transparency Test Score: We designed a question to assess participant's understanding of the robot's decision-making process. The question asks the participants to name the components of the robot that need repair. The components include the NBC sensors, audio and video processing units, etc. For the current study, only the audio/video processing units are faulty. Participants receive either 0 or 1 on this test.

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 in each mission (15). The value ranges from 0 to 100%.

Correct Decisions: This 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 in each mission (15). The value ranges from 0 to 100%.

5 Results

Data from 61 participants are included in the analysis (14 women, 39 men, $M_{age} = 19.2$ years, age range: 18-23 years). 2 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 actively involved in such missions.

We conducted a General Linear Model analysis with Repeated Measures and Bonferroni corrections, using explanation, embodiment, and promise to improve as within-subject factors, and trust, transparency, compliance, and correct decisions as dependent variables. Results show that explanation had a significant impact on trust ($F(1, 60) = 118.68$, $p < .0001$), transparency ($F(1, 60) = 33.82$, $p < .0001$), transparency test score ($F(1, 60) = 11.72$, $p = .001$), compliance ($F(1, 60) = 66.31$, $p < .0001$), and correct decisions made ($F(1, 60) = 83.90$, $p < .0001$). As shown in Table 1, participants reported a higher level of trust in the robot's ability when it offered explanations

on its decisions. Participants also reported that they felt that they understood the robot's decision-making process better when the robot offered explanations. Additionally, when the robot offered explanations, the human teammate made better decisions, as reflected in the percentage of correct decisions. The explanation also helped the human teammate calibrate when they should trust robot, as indicated by the combination of compliance rate and percentage of correct decisions. For each mission, 80% of the robot's decisions are correct (12 out of 15). We can see from Table 1 that when the robot offered no explanations, the participants over-trusted the robot (89.3%), resulting in poor decisions (69%). In contrast, when the robot offered explanations, the compliance rate (78.9%) is much closer to the robot's correctness rate (80%). On the other hand, when the robot offered no explanations, participants scored higher on the transparency test score, compared to when explanations were offered.

Table 1. The effect of explanation on trust, transparency, compliance, and correctness.

	No Explanation	**Confidence Explanation**
Transparency	2.75 (out of 7)	3.65 (out of 7)
Transparency Test Score	.414 (out of 1)	.275 (out of 1)
Trust	2.99 (out of 7)	5.07 (out of 7)
Compliance	89.3%	78.9%
Correct Decisions	69%	85.1%

The main effect of the promise to learn from mistakes was not statistically significant for any of the dependent variables. The robot's embodiment had a marginally significant effect on trust ($F(1, 60) = 3.64$, $p = .061$) and no significant main effect on the rest of the dependent variables. With a dog-like embodiment, participants reported a lower level of trust, compared to that reported for the machine-like robot embodiment ($M_{dog} = 3.96$, $M_{robot} = 4.10$).

There was a marginally significant interaction between the robot's promise to improve and its explanations on trust ($F(1, 60) = 3.85$, $p = .054$). As shown in Table 2, when the robot offered explanations, additional acknowledgment of error and promise to learn from it did not make much difference in the self-reported trust in the robot. However, when the robot did *not* offer explanations, acknowledging that an error was made and promising to improve did lead to higher self-reported trust by the participants.

Table 2. Interaction between the robot's explanations and acknowledgment on participants' trust.

	No Explanation	**Confidence Explanation**
No ACK.	2.86 (out of 7)	5.09 (out of 7)
With ACK.	3.13 (out of 7)	5.05 (out of 7)

The analysis of the main effect of embodiment shows that there was a marginally significant impact of the robot's embodiment on trust, as we originally hypothesized. The rationale of the hypothesis is that participants will carry their trust relationship with the real animal over to the animal-like robot. Such carry-over may last only a short period of time after the initial interaction, as over time, it will be overcome by

the actual behavior of the robot. We plotted the self-reported trust over the course of 8 missions between interactions of robots with different embodiments. As Figure 3 shows, this hypothesized decaying effect is indeed the case. Initially (i.e., during the first mission), the trust level differed significantly between the two robot embodiments. An ANOVA with explanation, embodiment, and acknowledgement as fix factors and self-reported trust right after the first mission as the dependent variable shows a significant main effect of the robot's embodiment ($M_{dog} = 3.58$, $M_{robot} = 4.39$, $F(1, 60) = 6.48$, $p = .014$) and the explanation offered ($M_none = 2.91$, $M_{confidence} = 5.00$, $F(1, 60) = 47.47$, $p < .001$). However, over time (in fact, starting from the second mission), the impact of the robot's embodiment is overtaken by the robot's behavior. And the difference in the level of trust in the robot is dominated by the difference in the robot's behavior, in this case, mainly the explanations offered by the robot (Figure 3). The aforementioned ANOVA tests on self-reported trust after subsequent missions indicated only a significant main effect of the robot's explanations. Additionally, we did not observe a significant impact of embodiment on the other dependent variables during the first mission or over the course of 8 missions.

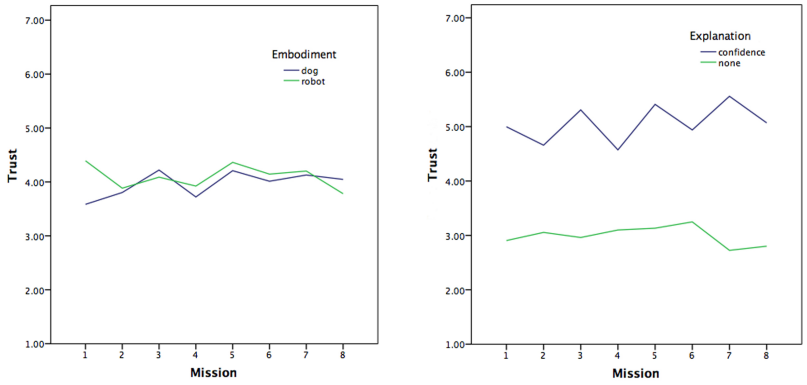

Fig. 3. Left: Self-reported trust in the robot between two different robot embodiments. Right: Self-reported trust in the robot between a robot offering confidence-level vs. no explanation .

6 Discussion

It is intriguing that explanation had a significant effect on transparency and self-reported trust. The explanation offered by the robot indicates only its confidence in its own assessment, not any information about what the assessment was based on or how the assessment was made. However, participants still reported that they felt that they understood the robot's decision-making process better when such an explanation was offered. Furthermore, such explanations do not indicate which component of the robot is faulty. This is reflected in the generally low scores on the transparency tests ($M = .344$). It may be somewhat surprising that when the robot offered explanations, participants actually scored lower on the transparency test. Perhaps, during these times, participants relied on the robot's explanation to guide their decisions, without thinking about where its recommendation came from. When the robot offered no explanations, the participants had to rely on experience with the robot to make future decisions, and, in the

process, tried to figure out what was wrong with the robot. Additionally, by offering explanations of its decisions, the robot helped its teammate calibrate a proper compliance rate (e.g., when and when not to follow its recommendations), resulting in better decision-making by the human teammate. Finally, without receiving explanations from the robot, the participants overused the robot, i.e., they followed the robot's recommendations even when the robot was wrong. Such participants also reported a lower level of trust in the robot, most likely due to the fact that over-relying on the robot during the mission resulted in poor decision-making.

Embodiment had an only marginally significant impact on self-reported trust. In particular, participants reported a lower level of trust in a robot with a dog-like appearance, compared to a robot with a machine-like appearance. This difference in trust levels did not translate to a difference in perceived transparency, compliance, or percentage of correct decisions. The direction in which the self-reported trust differed is contrary to our hypothesis. We hypothesized that one would trust a robot that looks like a dog more because of the existing trust relationship between humans and their best friend. The fact that participants trusted a machine-like robot more indicates that trust in the robot's ability could be context- and task-dependent. A machine-like robot may give the initial impression of a technological advantage offered by more sophisticated sensing equipment, which would be relevant to the reconnaissance task, compared to a dog-like robot. However, the effect of embodiment on self-reported trust is most pronounced at the beginning of the interaction. After interacting with the robot in the first mission, the robot's embodiment made no significant difference in the teammate's trust. The robot's behavior, particularly the explanations, which are highly relevant to the teammate's decision-making and team performance, became the deciding factor on transparency, trust, compliance, and teammate's decision-making for the following missions. This is congruent with the notion that trust is built based on past experience and first impressions based on appearance may not last.

Acknowledging a mistake and promising to learn from it had no significant effect on any of the dependent measures in the study. This is contrary to our hypothesis. Perhaps this is due to the fact that the participants interacted with each robot in only one mission. While the robot promised to update its decision-making algorithms to reflect the experience during the mission, the participants never got to witness any change after that mission. Making such acknowledgment had a significant interaction effect with the robot's explanations on self-reported trust. Particularly, when no explanations were offered, making such acknowledgment and promise can help restore some trust and perhaps instill some hope in the robot. However, when the robot offered explanations with its recommendations, such acknowledgement and promise did not make any significant impact. Perhaps the explanations alone were enough to steer the trust relationship.

We experimented with a robot that acknowledged its errors and promised to learn from them after the mission. However, it did not change its behavior within the mission, possibly accounting for the lack of any significant effect of such an acknowledgment/promise. Our robot's POMDP model can support such a dynamic behavior with some straightforward modification. In the next iteration of this study, we will expand the robot's POMDP model to include an explicit representation of the possibility of a sensor failure. The standard POMDP belief-update algorithms could then allow this

robot to decrease its confidence in its vision system after each false negative it receives [5]. In general, we can allow our robot to perform model-based *reinforcement learning* to update any aspect of its POMDP model [37]. Introducing the ability for the robot to change its decision-making model will no doubt raise new challenges for maintaining transparency and trust for human teammates.

7 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. Lewis, M., Sycara, K., Walker, P.: The role of trust in human-robot interaction. In Abbass, H.A., Scholz, J., Reid, D.J., eds.: Foundations of Trusted Autonomy. Springer-Verlag (2017)
2. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. Human factors **39**(2) (1997) 230–253
3. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1) (2004) 50–80
4. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10) (1992) 1243–1270
5. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial intelligence **101**(1) (1998) 99–134
6. Wang, N., Pynadath, D.V., Hill, S.G.: The impact of POMDP-generated explanations on trust and performance in human-robot teams. In: International Conference on Autonomous Agents and Multiagent Systems. (2016)
7. Schweitzer, M.E., Hershey, J.C., Bradlow, E.T.: Promises and lies: Restoring violated trust. Organizational behavior and human decision processes **101**(1) (2006) 1–19
8. Walters, M.L., Koay, K.L., Syrdal, D.S., Dautenhahn, K., Boekhorst, R.T.: Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials. In: AISB Symposium on New Frontiers in Human-Robot Interaction Convention. (2009) 136–143
9. Bruemmer, D.J., Marble, J.L., Dudenhoeffer, D.D.: Mutual initiative in human-machine teams. In: IEEE Conference on Human Factors and Power Plants, IEEE (2002) 7–22–7–30
10. Billings, D.R., Schaefer, K.E., Chen, J.Y., Kocsis, V., Barrera, M., Cook, J., Ferrer, M., Hancock, P.A.: Human-animal trust as an analog for human-robot trust: A review of current evidence. Technical Report ARL-TR-5949, Army Research Laboratory (2012)
11. Kerepesi, A., Kubinyi, E., Jonsson, G., Magnusson, M., Miklosi, A.: Behavioural comparison of human-animal (dog) and human-robot (AIBO) interactions. Behavioural processes **73**(1) (2006) 92–99
12. Melson, G.F., Kahn, P.H., Beck, A., Friedman, B., Roberts, T., Garrett, E., Gill, B.T.: Children's behavior toward and understanding of robotic and living dogs. Journal of Applied Developmental Psychology **30**(2) (2009) 92–102
13. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. Journal of Human-Computer Studies **58**(6) (2003) 697–718
14. Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. In: Second generation expert systems. Springer (1993) 543–585

15. Elizalde, F., Sucar, L.E., Luque, M., Diez, J., Reyes, A.: Policy explanation in factored markov decision processes. In: European Workshop on Probabilistic Graphical Models. (2008) 97–104
16. Visschers, V.H.M., Meertens, R.M., Passchier, W.W.F., De Vries, N.N.K.: Probability information in risk communication: a review of the research literature. Risk Analysis **29**(2) (2009) 267–287
17. Hendrickx, L., Vlek, C., Oppewal, H.: Relative importance of scenario information and frequency information in the judgment of risk. Acta Psychologica **72**(1) (1989) 41–63
18. Waters, E.A., Weinstein, N.D., Colditz, G.A., Emmons, K.: Formats for improving risk communication in medical tradeoff decisions. Journal of health communication **11**(2) (2006) 167–182
19. Matarić, M.J.: Reinforcement learning in the multi-robot domain. Autonomous Robots **4**(1) (1997) 73–83
20. Smart, W.D., Kaelbling, L.P.: Effective reinforcement learning for mobile robots. In: IEEE International Conference on Robotics and Automation. Volume 4., IEEE (2002) 3404–3410
21. Lewicki, R.J.: Trust, trust development, and trust repair. In Deutsch, M., Coleman, P.T., Marcus, E.C., eds.: The handbook of conflict resolution: Theory and practice. Wiley Publishing (2006) 92–119
22. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. In: International Conference on Social Robotics, Springer (2015) 574–583
23. Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: Comparing automatically generated explanations. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction, Piscataway, NJ, USA, IEEE Press (2016) 109–116
24. Wang, N., Pynadath, D.V., Hill, S.G.: Building trust in a human-robot team. In: Interservice/Industry Training, Simulation and Education Conference. (2015)
25. Rovira, E., Cross, A., Leitch, E., Bonaceto, C.: Displaying contextual information reduces the costs of imperfect decision automation in rapid retasking of isr assets. Human factors **56**(6) (2014) 1036–1049
26. Wickens, C.D., Dixon, S.R.: The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science **8**(3) (2007) 201–212
27. Pop, V.L., Shrewsbury, A., Durso, F.T.: Individual differences in the calibration of trust in automation. Human factors **57**(4) (2015) 545–556
28. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Academy of management review **20**(3) (1995) 709–734
29. McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and validating trust measures for e-commerce: An integrative typology. Information systems research **13**(3) (2002) 334–359
30. McShane, S.L.: Propensity to trust scale (2014)
31. Ross, J.M.: Moderators of trust and reliance across multiple decision aids. ProQuest (2008)
32. Syrdal, D.S., Dautenhahn, K., Koay, K.L., Walters, M.L.: 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)
33. Greco, V., Roger, D.: Coping with uncertainty: The construction and validation of a new measure. Personality and individual differences **31**(4) (2001) 519–534
34. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. Advances in psychology **52** (1988) 139–183
35. Taylor, R.M.: Situational awareness rating technique(sart): The development of a tool for aircrew systems design. Situational Awareness in Aerospace Operations (1990)
36. Mayer, R.C., Davis, J.H.: The effect of the performance appraisal system on trust for management: A field quasi-experiment. Journal of applied psychology **84**(1) (1999) 123
37. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of artificial intelligence research **4** (1996) 237–285