Comparing Two Automatically Generated Explanations on the Perception of a Robot Teammate

Ning Wang
David V. Pynadath
Institute for Creative Technologies
University of Southern California
nwang@ict.usc.edu,pynadath@usc.edu

Michael J. Barnes
Susan G. Hill
US Army Research Laboratory
michael.j.barnes.civ@mail.mil,susan.g.hill.civ@mail.mil

ABSTRACT

Trust is critical to the success of human-robot interaction (HRI). Research has shown that people will more accurately trust a robot if they have a more accurate understanding of its decision-making process. Recent research has shown promise in calibrating human-agent trust by automatically generating explanations of decision-making process such as POMDP-based ones. In this paper, we compare two automatically generated explanations, one with quantitative information on uncertainty and one based on sensor observations, and study the impact of such explanations on perception of a robot in human-robot team.

1 INTRODUCTION

Trust is critical to the success of human-robot interaction (HRI) [7]. Research has shown that people will more accurately trust an autonomous system, like a robot, if they have a more accurate understanding of its decision-making process [6]. The Partially Observable Markov Decision Process (POMDP) is one such decision-making process, providing a framework for optimized decision making that is commonly used by robots, agents, and other autonomous systems [5]. While the computations required by POMDP algorithms typically obfuscate the decision-making process from people, recent research has shown promise in calibrating human-agent trust by automatically generating explanations of POMDP-based decisions [11].

In this paper, we build on our previous work in automatically generated explanations and seek a deeper understanding of the impact of such explanations on human perceptions of a robot. We specifically focus on comparing two types of explanations: one that provides quantitative information on uncertainty and one that provides detailed information about sensor readings. We extend our previous discussions based on measures of trust and team performance to consider human perceptions of the robot, such as predictability, a factor critical to trust in HRI. We use a testbed that stimulates trust behaviors when interacting with a simulated robot. By comparing online participants with robots using different explanations, the analysis presented here can have useful implications for the design of explanation-mechanisms for robots of the future.

2 RELATED WORK

Explanations have shown to contribute to people's understanding of a robot's decisions in a way that provides transparency and improves trust [1]. Our goal is to create an automated, domain-independent method for generating explanations that has the same impact as the manually crafted explanations used in prior work. Recent work on generating explanations based on Markov Decision Problems (MDPs) [2, 11] has shown promise as to the potential success of applying a general-purpose explanation on top of an agent's decision-making process.

There are a variety of studies that measure the impact of forms of explanation on people's perceptions of risks and uncertainties when making decisions. A survey of these studies across multiple domains indicates that "people prefer numerical information for its accuracy but use a verbal statement to express a probability to others." [9]. On the other hand, one study 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 [4]. 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 [13]. 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.

3 HRI TESTBED

We carry out our investigation in the context of an online HRI testbed, described in detail in [10]. For the study discussed here, we configured the testbed to implement a scenario in which a human teammate works with a robot to carry out three reconnaissance missions that require the human teammate to search buildings in a foreign 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 fatally injured if not wearing the protective gear. In such a case, the team incurs a 3-minute time penalty. However, it takes time to put on and take off protective gear (20 seconds each). If the human fails to search all the buildings in town within the time limit, the mission is a failure. So the human is incentivized to consider the robot's findings before deciding how to enter the building. Details on how the task is modeled as an Partially Observable Markov Decision Process (POMDP) are discussed in [12].

HRI'18, March 2018, Chicago, IL, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 EVALUATION

We used the online testbed to conduct an evaluation study on how the robot's explanation impacted trust and team performance. Details of the study methodology can be found in [12]. In this paper, we focus on the comparing two automatically generated explanations, confidence-level explanation and observation explanation. Both explanation include a robot's decision from its scouting report (e.g., "I have finished surveying the doctor's office. I think the place is safe.") The confidence-level explanations augment the decision message with additional information about the robot's uncertainty in its decision (e.g., "I am 78% confident about this assessment."). The observation explanations instead augment the decision message with non-numeric information about the robot's sensing capability (e.g., "My sensors have not detected any NBC weapons in here. From the image captured by my camera, I have not detected any armed gunmen in the cafe. My microphone picked up a friendly conversation.") These explanations will thus potentially help the robot's teammate understand which sensors are working correctly and which ones are not.

Previous analyses have shown that both type of explanations fostered the robot's trust relationship with its human teammate and improved transparency communication and team performance, compared to when no explanations were offered. And no significant differences were found between these two type of explanations on self-reported trust and team performance ([12]. This paper focuses on participants' perception of the robot, measured using items from [8] and items developed by the researchers (e.g., items regarding the robot's ability to be aware of its limitations, which can particularly be influenced by confidence level explanations).

5 RESULTS

The trust in robots scale by [8] includes 40 items, presented in 10-point Likert scale. Independent-samples t-tests were conducted to compare the perception of the robot for participants who received confidence-level explanations and sensor-observation explanations. There were significant differences on perceptions that the robot did "openly communicate" ($M_{conf} = 8.77$, $M_{obs} = 7.38$, $p = .0262$), "processes adequate decision-making ability" ($M_{conf} = 8.67$, $M_{obs} = 7.88$, $p = .0447$) and "predictable" ($M_{conf} = 7.28$, $M_{obs} = 8.30$, $p = .0234$). No significant difference was found on the other items from the scale.

Independent-samples t-test was also conducted to compare the perception of if "The robot is aware of its own limitations" (7-point Likert scale). Results show that participants rated the robot that provided confidence-level explanation significantly higher than the robot that provided sensor-observation explanation ($M_{conf} = 5.06$, $M_{obs} = 4.02$, $p = .0148$).

6 DISCUSSION

In this paper, we compared the impact of two different automatically generated explanations on the perception of the robot in in human-robot teams. The results show that a robot with explanations that contain numeric information of uncertainty was perceived to be more openly communicative with more adequate decision-making ability but less predictable, compared to a robot with explanations that present its sensor readings (and in non-numeric form). This result is somewhat inline with the result on the perception of the robot's ability to be self-aware of its limitations — a robot with confidence-level explanations is perceived with higher on this ability. The confidence-level explanation contains information of the robot's assessment of the uncertainty of its own decisions. Communication of this information indicates that the robot is aware of its shortcomings, yet still willing to relay that to its teammate. Perhaps the confidence-level explanation served as an easy-to-parse decision heuristic for the human teammate, while presenting sensor observations did not necessarily highlight its ability to make decisions. Thus, a robot with confidence-level explanations was perceived to be a better decision-maker. Interestingly, however, a robot with confidence-level explanations is perceived to be less predictable. This may be due to the fact that such information does not explain why the robot's confidence level changes.

In our previous work, we did not find any significant difference between these two types of explanations on perceived trust and transparency, compliance with the robot's recommendations, and team performance. The results presented here indicate that not only are these two explanations not created equal, but they may have impact on human-robot teaming and human teammates' behaviors. For example, those who are less adept with uncertainty, emotionally and cognitively [3], may not work well with a robot that they perceive to be not as predictable. Further research is under way to study the role of individual differences in the design of automatically generated explanations to build trust in human-robot teams.

REFERENCES

[1] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.

[2] Francisco Elizalde, L. Enrique Sucar, Manuel Luque, J. Diez, and Alberto Reyes. 2008. Policy explanation in factored Markov decision processes. In *Proceedings of the European Workshop on Probabilistic Graphical Models.* 97–104.

[3] Veronica Greco and Derek Roger. 2001. Coping with uncertainty: The construction and validation of a new measure. *Personality and individual differences* 31, 4 (2001), 519–534.

[4] Laurie Hendrickx, Charles Vlek, and Harmen Oppewal. 1989. Relative importance of scenario information and frequency information in the judgment of risk. *Acta Psychologica* 72, 1 (1989), 41–63.

[5] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1 (1998), 99–134.

[6] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[7] Michael Lewis, Katia Sycara, and Phillip Walker. 2017. The Role of Trust in Human-Robot Interaction. In *Foundations of Trusted Autonomy*, Hussein A. Abbass, Jason Scholz, and Darryn J Reid (Eds.). Springer-Verlag.

[8] Kristin E Schaefer. 2013. *The perception and measurement of human-robot trust.* Ph.D. Dissertation. University of Central Florida Orlando, Florida.

[9] Vivianne H. M. Visschers, Ree M. Meertens, Wim W. F. Passchier, and Nanne N. K. De Vries. 2009. Probability information in risk communication: a review of the research literature. *Risk Analysis* 29, 2 (2009), 267–287.

[10] Ning Wang, David V. Pynadath, and Susan G. Hill. 2015. Building Trust In a Human-Robot Team. In *Interservice/Industry Training, Simulation and Education Conference.*

[11] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *International Conference on Autonomous Agents and Multiagent Systems.*

[12] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust Calibration Within a Human-Robot Team: Comparing Automatically Generated Explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16).* IEEE Press, Piscataway, NJ, USA, 109–116.

[13] Erika A. Waters, Neil D. Weinstein, Graham A. Colditz, and Karen Emmons. 2006. Formats for improving risk communication in medical tradeoff decisions. *Journal of health communication* 11, 2 (2006), 167–182.