The Effect of Perceived Involvement on Trust in Human-Robot Interaction

Daniel Ullman and Bertram Malle
Cognitive, Linguistic, and Psychological Sciences
Brown University
Providence, RI, USA
{daniel_ullman, bertram_malle}@brown.edu

***Abstract*—Trust serves as a powerful social capacity that can influence the course of a relationship, either spurring a willingness or refusal of one agent to interact with another. As we attempt to build increasingly complex and useful social robots, we must consider what factors will engender such trust and thus benefit human-robot interaction. In this paper we describe a line of inquiry that is investigating how a person's perceived involvement in helping a robot recover from failure affects the person's trust in the robot and in its future actions. We posit that a person's active involvement with a robot, compared with passive observation, will lead to greater trust in the robot.**

Keywords—trust; agency; social robotics; human-robot interaction

I. Introduction

Social robotics—the development of robots that can excel in social contexts involving human-robot interaction (HRI)—has the potential to enhance the daily lives of human beings. But for HRI to offer any of the benefits that social robots promise, a person must want to engage with a robot.

Robots are going to fail. While robots are often able to perform certain tasks with a high rate of success, they are still vulnerable to error. This research is motivated by the concept of failure and, in particular, by the need to recover from failure. Robots, like humans, must recover from failure. When a human makes a mistake, one of three things happens: In the first case, the person fails to correct for the error and continues to make the same mistake; the person does not learn. In the second case, the person self-corrects for the error; the person learns without external help. In the third case, the person corrects for the error based on feedback from some outside source; the person learns with external help. These distinctions are applicable to robots as well. The present study contrasts robots that appear to self-correct autonomously with robots that appear to self-correct with external help.

To pursue and regulate goals is to demonstrate agency, where beliefs about one's capacity to exercise control in the outside world are central to notions of human agency [1]. Failure often prompts a loss of trust in an agent. But trust can be restored if an agent, specifically a robot, recovers from failure by way of learning and improvement [2]. In this project we examine how people think and feel about a robot that appears to correct for error in different ways; in particular, we investigate whether people attribute greater trust to a robot, such that they are more willing to interact with it, when they believe they can influence the robot.

In this experiment we manipulate a person's perceived involvement with a robot's actions and measure the person's resulting willingness to interact with similarly designed robots in the future. The manipulation of *perceived involvement* holds constant the actual effect of the participant's involvement with the robot; the participant's actions appear to influence the robot's actions, but in reality do not have any causal effect. We posit the following hypotheses, whereby H2 represents a moderator of H1:

> **H1:** Participants will perceive a robot to be more trustworthy when they believe they can influence the robot.
>
> **H2:** Participants will exhibit increased trust in a robot that they believe they can influence only in scenarios where the robot's actions directly affect a person.

As robots are increasingly utilized in interactive social and socially assistive settings, we must further explore expectations about robots in these interactive relationships. This investigation is warranted as part of a broader inquiry into how people think about robots in moral contexts, given the potential for robots to interact with people in vulnerable settings [3].

II. Related Work

Successful relationships are typically founded on trust, which is integral to continued interaction between two agents [4]. It is therefore imperative that we investigate the nature of trust in the context of HRI. As most benefits offered by social robots rely on repeated interactions with a person, a robot must not only generate trust but also maintain that trust to enable a beneficial relationship. Trust seems to affect the extent to which a person is willing to allow a robot to act autonomously [5]. Definitions of trust vary, but typically they involve a common element where one agent has a belief that another agent will act in a certain way—trust therefore relies on some set of expectations about an agent's actions [6].

In HRI, various forms of human involvement with a robot influence trust of the robot. We examine here whether involvement itself, abstracted from specific forms of interaction, has an impact on trust; is the mere illusion of causal impact on an agent, and the agent's perceived

responsiveness to this impact, sufficient to generate trust? Previous work has studied trust in contexts involving cheating robots. In one study participants interacted directly with a robot, where participants themselves were affected by the robot's cheat behavior; participants lost trust in the robot even when a few cheat occurrences were surrounded by a majority of honest interactions [7]. In another study participants observed either a robot or a person in a cheating scenario, where participants themselves were not affected by the cheating; participants evaluated humans and robots differently in contexts involving trust, ascribing different levels of intelligence and trustworthiness to each type of agent in identical scenarios [8]. While previous research has examined trust in contexts containing varying levels of involvement, there has yet to be a systematic study of how manipulating level of involvement affects attributions of trust to a robot.

III. Method

In this experiment the minimalist educational robot Thymio attempts to complete a simple task of moving from a starting location to an end location, navigating an obstacle in between. The experimental design utilizes the self-propelled movement of the robot as a basic cue to agency for the robot [9]. This obstacle avoidance task was chosen for its simplicity in order to avoid more complex social constructs.

All participants take part in one of three conditions of perceived involvement, with participants randomly assigned to conditions. Participants are told that the robot has one of the three following types of software; each piece of software enables it to generate and execute a possible path of motion, and it can recognize if it encounters an obstacle and fails to reach its goal. In each condition participants first observe the robot as it fails to achieve the goal state, bumping into the obstacle and stopping its movement. Then participants experience the manipulation of perceived involvement:

Autonomous Involvement. Participants are told the robot's software generates and executes an alternate path. Participants watch the robot complete its task.

Experimenter Involvement. Participants are told the robot's software waits for a button press from the experimenter to tell it to generate and execute an alternate path. Participants watch the experimenter press the robot's button, and watch the robot complete its task.

Participant Involvement. Participants are told the robot's software waits for a button press from the participant to tell it to generate and execute an alternate path. Participants press the robot's button, and watch the robot complete its task.

In all conditions the robot executes the alternate path after a timeout of 10 seconds; the button press is merely an instance of perceived control, and in fact does not affect the actions of the robot. After observing the robot complete its task in the assigned condition, participants are presented with a set of Likert items to evaluate perceptions of trustworthiness, agency, and likability; these data will be used to test hypothesis H1. Then participants complete a final task in which they are presented with a set of vignettes that describe possible contexts where a robot could be used. Participants complete Likert items on how well suited a robot with the observed type of software would be for the scenario, and rate to what extent they would trust a robot in the particular scenario. These scenarios, presented in random order, fit into two categories: scenarios where the robot's actions directly affect a person, and scenarios where the robot's actions do not directly affect a person. These data will be used to test hypothesis H2.

We are running 60 participants, with 20 participants in each condition. We will conduct a one-way analysis of variance (ANOVA) to test the discussed hypotheses.

IV. Conclusion

HRI, and in particular social robotics, requires a relationship between two agents. To attain many of the benefits offered by social robots, we must understand what motivates people to engage with them. Trust is one of the core components of social interaction, and it is therefore imperative to understand what engenders trust in interactions with robots. We are investigating whether a person's perceived involvement in a robot's recovery from failure increases trust in a robot. As we better understand these human responses, we will be able to design robots that increasingly support humans through HRI.

Acknowledgments

This material is based upon work supported by the Office of Naval Research (MURI grant #N00014-14-1-0144, Moral Competence in Computational Architectures for Robots: Foundations, Implementations, and Demonstrations). Daniel Ullman is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

[1] A. Bandura (1989). Human agency in social cognitive theory. *American Psychologist*, *44*, 1175-1184.

[2] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, & H. Yanco (2013). Impact of robot failures and feedback on real-time trust. *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. Tokyo, Japan, March 3-6.

[3] B. F. Malle, & M. Scheutz (2014). Moral competence in social robots. *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics 2014*. Chicago, Illinois, May 23-24.

[4] D. Gambetta (1988). *Trust: Making and breaking cooperative relations*. Oxford, UK: Basil Blackwell.

[5] M. Desai, K. Stubbs, A. Steinfeld, & H. Yanco (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. *Proceedings of the AISB Convention: New Frontiers in Human-Robot Interaction*. Edinburgh, Scotland, April 6-9.

[6] J. Rotter (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*, 651-656.

[7] A. Litoiu, D. Ullman, J. Kim, & B. Scassellati (2015). Evidence that robots trigger a cheating detector in humans. *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*. Portland, Oregon, March 2-5.

[8] D. Ullman, I. Leite, J. Phillips, J. Kim-Cohen, & B. Scassellati (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Quebec City, Canada, July 23-26.

[9] D. Premack (1990). The infant's theory of self-propelled objects. *Cognition*, *36*, 1-16.