Overtrust of Robots in High-Risk Scenarios

Jin Xu

Human-Automation Systems Lab (HumAnS Lab), Georgia Institute of Technology, Atlanta, GA, USA
jin.xu@gatech.edu

Introduction

From personal robot assistant to self-driving vehicles, artificial intelligence (AI) is the backbone underlying millions of future advanced applications. As robots become increasingly pervasive in daily life, it is expected that robots will augment human laborers in many domains in the near future. When robots are deployed in the real world, the underlying assumption is that they are capable of accomplishing their given tasks. However, researchers have shown that robots made mistakes, and in several cases, humans tend to overtrust robotic systems (Abney 2017; Borenstein et al. 2017; Robinette, Howard, and Wagner 2017). Overtrust of a robot happens in scenarios where "(1) a person accepts risk because that person believes the robot can perform a function that it cannot or (2) the person accepts too much risk because the expectation is that the system will mitigate the risk." (Abney 2017). In particular, we are interested in two emerging domains where an appropriate amount of trust is a minimal requirement and overtrust could cause harm: 1) healthcare scenarios and 2) self-driving car (i.e. autonomous driving) scenarios. Both healthcare and autonomous driving scenarios often involve high risks, and the negative outcomes could be detrimental to the user. The objective of our research focuses on 1) investigating the causes that contribute to human overtrust of these robots systems 2) developing a behavior-based computational model to predict overtrust, and 3) developing techniques to mitigate outcomes caused by the overtrust.

Trust in Healthcare Scenarios

Embodied AI agents, or interactive robots, have been largely utilized in a variety of healthcare scenarios and will increasingly occupy the healthcare realm (Kiesler et al. 2008; Pak et al. 2012; Brown, García-Vergara, and Howard 2015). As a precursor for successful human-human interactions, trust plays a critical role in maintaining a positive patient-doctor relationship (Goold and Lipkin 1999). Several studies have raised concerns regarding overtrust in utilizing embodied agents in the healthcare realm by physicians, caregivers, and children (Borenstein, Wagner, and Howard 2017; Yamagishi 2011).

We started investigating human-robot trust in a typical healthcare scenario involving the disclosure of personal information (Xu and Howard 2017). The study found that humans will trust a socially interactive robot and disclose to them personal information, even when interacting with a faulty robot, suggesting a potential overtrust of robots. In order to get further insights from a realistic healthcare scenario, we integrated a humanoid robot with an upper-body rehabilitation therapy game to function in the role of therapist (Lee, Xu, and Howard 2017), and conducted a comparison study examining outcomes when using a robotic agent versus a human agent. This study found that the robot therapist improved participant's motor performance faster than a human therapist, and trust was equivalent between the two conditions. This study provides preliminary evidence that humans may blindly follow a robot's guidance without evaluation the potential risks. Future direction of this research includes developing techniques that extract trust from the user through the emotional feedback of the robot and developing methods to mitigate overtrust using emotional feedback.

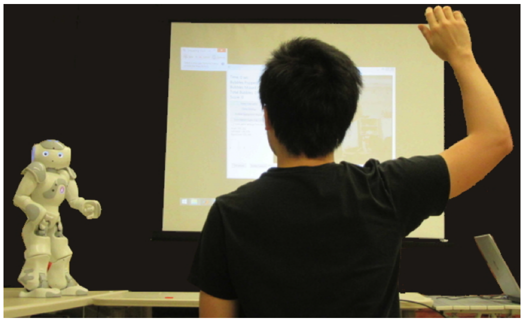

Figure 1: Example of a participant interacted with the NAO robot while playing SuperPop VR therapy game

Trust in Autonomous Vehicles

Recently, several technology companies have outlined their ambitious plans to bring fully autonomous vehicles to the market in the next 5 years (Muoio 2017). Although autonomous vehicles have the potential to considerably reduce traffic accidents and improve safety, concerns remain regarding imperfect performance of self-driving vehicles as compared to actual human drivers. In the interaction involving an autonomous vehicle, the trust problem is not just related to the passenger but also related to other parties that share the same road with the autonomous vehicles (e.g. manually driven vehicle or pedestrians). Several studies have found that the public tends to overtrust robots by assuming robots have capabilities beyond their actual ability (Abney 2017; Borenstein et al. 2017; Robinette, Howard, and Wagner 2017). This phenomenon of overtrust is also expected to exist in self-driving scenarios, and we seek to investigate this problem before deploying them on the road.

This work aims to investigate human reaction and further understand human trust towards autonomous vehicles they encounter on the road. We have developed a virtual environment where participants interact with either a self-driving car or a manually driven car. Various assets (e.g. vehicles, buildings, intersections and traffic signs) are integrated into the environment to enhance the realism of the simulation. We have begun to investigate human-robot trust with a pilot study to examine differences in driver intersection behavior as they encounter manually and autonomously driven cars. This study aims to answer the following questions:

Q1: Do human driver reactions during stop sign intersection encounters with manually-driven cars differ from those exhibited during stop sign intersection encounters with autonomous cars?

Q2: Does overall driver thoughts and behavior differ when encountering manually-driven vehicles versus autonomous vehicles?

Future work of this research will focus on 1) expanding the experiments to include additional scenarios 2) incorporating physiological measurements of participants such as heart rate and eye tracking 3) developing a computational model to predict user's trust based on their behaviors and 4) developing techniques to prevent overtrust and/or mitigate risk caused by overtrust in autonomous driving scenarios.

Future Work

As robots gradually penetrate our home and workforces, understanding human-robot trust becomes increasingly important. We must address the potential risk of overtrust, and develop a set of methods to mitigate that risk before deploying them into the real world.

Studies in previous sections highlight important implications and directions for future works. The most obvious direction is to identify the behavioral and emotional factor that contribute to overtrust in a certain scenario. For example, a driver's facial expression may indicate their level of trust during self-driving scenarios. A fusion of multiple sensors will be used to assess user's behavior and emotional states. Next, we will develop a computational model to predict user's trust based on sensor measurements for a certain scenario. In addition, we will develop techniques to help to mitigate overtrust and further conduct experiments in both healthcare and self-driving scenarios.

References

Abney, K. 2017. Robot Ethics 2. 0: New Challenges in Philosophy, Law, and Society. Oxford University Press. pp. 127.

Borenstein, J.; Wagner, A.; and Howard, A. M. 2017. A case study in caregiver overtrust of pediatric healthcare robots. RSS Workshop on Morality and Social Trust in Autonomous Robots, Cambridge, MA.

Brown, L.; Garcia-Vergara, S.; and Howard, A. M. 2015. Evaluating the effect of robot feedback on motor skill performance in therapy games. Systems, Man, and Cybernetics (SMC), IEEE International Conference on, pp. 1060-1065.

Goold, S. D.; and Lipkin, M. 1999. The Doctor-Patient Relationship: Challenges, Opportunities, and Strategies. Journal of General Internal Medicine, 14(Suppl 1), S26-S33.

Kiesler, S.; Powers, A.; Fussell, S. R.; and Torrey, C. 2008. Anthropomorphic interactions with a robot and robot–like agent. Social Cognition, 26(2), 169-181.

Lee, B.; Xu, J.; and Howard, A. M. 2017. Does Appearance Matter? Validating Engagement in Therapy Protocols with Socially Interactive Humanoid Robots, IEEE Symposium Series on Computational Intelligence, Honolulu, HI, Nov. 2017.

Muoio, D 2017. 19 companies racing to put self-driving cars on the road by 2021. Business Insider. Retrieved 27 November 2017, from http://www.businessinsider.com/companies-making-driverless-cars-by-2020-2016-10

Pak, R.; Fink, N.; Price, M.; Bass, B.; and Sturre, L. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics, 55(9), 1059-1072.

Robinette, P.; Li, W.; Allen, R.; Howard, A. M.; and Wagner, A. R. 2016. Overtrust of robots in emergency evacuation scenarios. Paper presented at the Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on.

Xu, J.; and Howard, A. 2017. Pilot Study For Examining Human-Robot Trust In Healthcare Interventions Involving Sensitive Personal Information, Rehabilitation Eng. and Technology Society of North America (RESNA) Annual Conference, New Orleans, LA, June 2017.

Yamagishi, T. 2001. Trust as a form of social intelligence. In K. S. Cook (Ed.), Russell Sage foundation series on trust, Vol. 2. Trust in society (pp. 121-147). New York: Russell Sage Foundation.