Modeling Human-Robot Trust in Emergencies

Paul Robinette[1,2], Alan R. Wagner[2], and Ayanna M. Howard[1]

[1]School of Electrical and Computer Engineering
Georgia Institute of Technology
[2]Georgia Tech Research Institute
Atlanta, GA, USA 30332
probinette3@gatech.edu, alan.wagner@gtri.gatech.edu, ayanna.howard@ece.gatech.edu

Abstract

Modeling human trust decisions is a notoriously difficult problem. We focus on decisions where a victim must decide whether to trust a robot in an emergency situation and outline the necessary inputs to model this decision. These inputs can each be represented as an outcome matrix and combined using a weighted sum. Calibrating these weights can be accomplished through the use of internet surveys.

Introduction

In previous work, we have identified numerous situations that benefit from human-robot interaction in emergency domains (Robinette and Howard, 2011). Most of our work has focused on robots guiding evacuees out of a building during an emergency, but these same robots and algorithms could be applied to many other emergency situations. Thus far, we have assumed that these robots can be programmed to respond to any situation in a way that is guaranteed to increase survivability. We have also assumed that we can design the exterior of the robot in such a way that evacuees and other victims of an emergency are sure to follow the robot's directions. Unfortunately, it is very likely that no one robot design can accomplish this. Even if it could, the robot must be able to understand the human victim's motivations and respond accordingly to provide the best aid possible for that particular situation.

Outcome matrices are one way to represent two individuals' motivations in an interaction. Previous work has shown that these matrices can be created and modified by robots during interactions with humans (Wagner, 2009). We have previously used outcome matrices to enumerate a number of situations in which an evacuee would not follow a robot during an emergency (Robinette et al., 2013). These situations all assume that the evacuee has some personal reason to avoid following the robot, but what if the victim is simply afraid of the robot itself? What if the victim does not believe that the situation requires robotic guidance? The robot must be able to understand why a victim would or would not accept guidance in terms of personal motivations as well as the victim's perceived risk of the situation. For the emergency domain, we consider risk in terms of the situation as well as the agent or robot. We further divide agent risk into the perceived risk caused by the appearance of the agent and the perceived risk caused by the actions of the agent. Ultimately, this model of the perceived risk of the victim produces a measure of whether the victim will trust the robot or not.

Defining Trust

Wagner defines trust as "a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustee has put its outcomes at risk" (Wagner and Arkin, 2011). He also denotes four conditions for trust:

1. The trustee does not act before the trustor.
2. The outcome received by the trustor depends on the actions of the trustee if and only if the trustor selects the trusting action.
3. The trustor's outcome must not depend on the action of the trustee when selecting the untrusting action.
4. The value of fulfilled trust is greater than the value of not trusting at all, is greater than the value of having one's trust broken.

This definition of trust is particularly appealing for emergency situations because it directly deals with risk. Robots in an emergency situation are not attempting to reward humans for their compliance with directions; they are trying to mitigate risk to human life. Furthermore, in the emergency domain we can assume that the perceived risk of trusting the robot will be inversely proportional to the value of the interaction stated in the definition above.

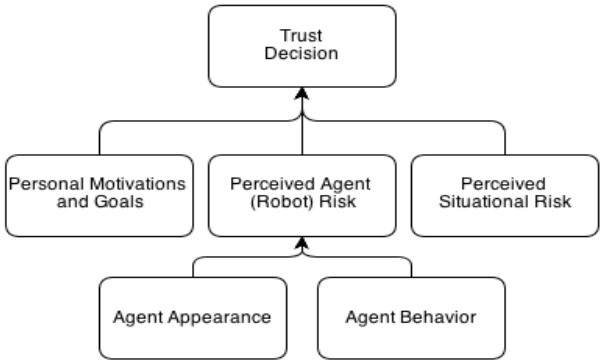

Figure 1: Diagram of a victim's decision to trust (or not) a robot's guidance in an emergency

Measuring Trust

Risk, as used in the definition of trust above and applied to the emergency domain, can be interpreted as a combination of situational risk and agent risk (Figure 1). Situational risk is the amount of danger that the victim perceives in the environment around him. Triggered fire alarms and the presence of smoke would increase the risk in a fire emergency. The sound of gunshots would increase the risk in an active shooter scenario. Very little risk might be perceived if there is no visual or audio indication of an emergency. Increased risk in the environment should generally increase the likelihood that the victim will follow the robot's directions in most situations.

Agent risk is considered in terms of both the agent's behavior and appearance. In this case, the agent is the robot trustee attempting to help the victim. In (Robinette and Howard, 2011) we have outlined the basic requirements for the appearance of an effective evacuation guidance robot. Other related work that has focused on the physical appearance and behavior of emergency response robots can also be useful here (Bethel and Murphy, 2008; Shell and Mataric, 2005). Following these guidelines will help to increase trust in the robot. In (Robinette et al., 2013), we examined which actions the robot could do to increase trust in situations where evacuees have personal reasons to disregard its directions. Many behaviors would increase the perceived risk of following the robot, such as the robot making the obvious error of colliding with an obstacle. There may be a perceived risk of the robot itself harming the victim. An increased perceived risk of following the guidance of the robot will generally lower the trust in the robot.

Each of the risks above can be represented as its own outcome matrix. Thus, a robot can represent an individual victim's motivations as a perceived situational risk outcome matrix, a perceived agent appearance risk outcome matrix and a perceived agent behavior risk outcome matrix. It is possible, and even likely, that these three matrixes will present conflicting information to the robot. If the robot can determine a rough importance measure for each matrix then it can perform a weighted sum of all the matrices and calculate how likely it is that the victim will accept the guidance of the robot.

Calibrating Trust Measures

Determining how a victim will respond to a robot sent to render assistance is difficult. Some information can be found by performing surveys. Outside of the trust domain, we have performed surveys to determine how workers on Amazon's Mechanical Turk service perceive guidance instructions from an emergency robot. For a small payment ($0.50 per survey), workers were asked to watch short videos of four different instructions and give their perception of each. At its peak response rate, 120 surveys from unique workers were submitted in approximately two hours. This same technique can be applied to the trust domain by showing short videos of a robot performing a task or providing assistance in environments with varying degrees of risk. The participant can then be asked either how much trust he or she has in the robot or what action he or she would perform after observing this robot. By using online surveys, we can determine how trust dynamics change with different robots in different circumstances.

Acknowledgements

Portions of this work were funded by award #FA95501310169 from the Air Force Office of Sponsored Research. References

References

Bethel, C. L. and Murphy, R. R. 2008. Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *IEEE Transactions on Systems, Man, And Cybernetics Part C*, 38(1):83–92.

Robinette, P., and Howard, A. 2011. Emergency evacuation robot design. In *ANS EPRRSD - 13th Robotics & Remote Systems for Hazardous Environments and 11th Emergency Preparedness & Response.*

Robinette, P., A. R. Wagner, and A. M. Howard. Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. *2013 AAAI Spring Symposium Series*.

Shell, D. A.; and Mataric, M. J. 2005. Insights toward robot-assisted evacuation. *Advanced Robotics*, 19(8):797–818.

Wagner, A. R. 2009. Creating and using matrix representations of social interaction. *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009.

Wagner, A. R., and Arkin, R. C. 2011. Recognizing situations that demand trust. *RO-MAN, 2011 IEEE*.