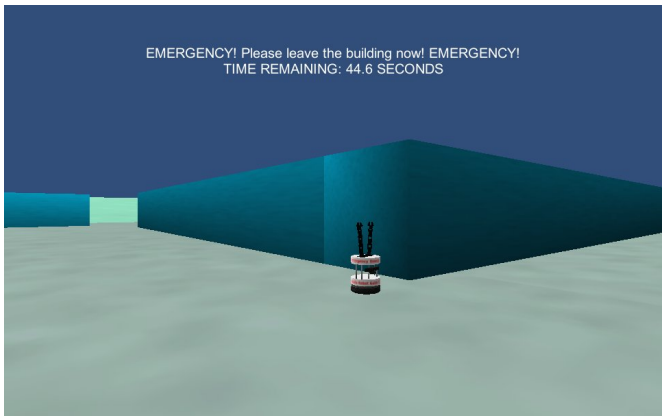
A conceptual framework for modeling human-robot trust

6 December 2018, by Ingrid Fadelli



The image depicts a maze navigation experiment that was described as an emergency situation. The participants were told that their task was to act as if they were in an emergency evacuation and had to find an exit within 30 seconds in order to survive. Credit: Wagner et al.

Researchers at Pennsylvania State University, MIT and Georgia Institute of Technology have recently developed a conceptual framework to model the human-robot trust phenomenon. Their framework, outlined in a paper published on ACM Digital Library, uses computational representations inspired by game theory to represent trust, a notion defined using theory in social psychology.

Trust plays a key role in interpersonal interactions, both in professional and personal settings. When a person trusts another, they might decide to do something that puts them at considerable risk, holding the belief that the actions of the other will somewhat mitigate this risk.

When it comes to defining trust, many researchers agree that it entails a situation in which an individual is vulnerable and this vulnerability rests with the actions, motivations, or behaviors of another. Alan R. Wagner, Paul Robinette, and Ayanna Howard, the three researchers behind the recent study, wanted to create a framework that could be used to model interpersonal trust between humans and robots.

"We are interested in developing robots that know when to trust humans, understand what situations require trust and attract the right amount of trust from humans," Robinette explained. "Initially, we were motivated by research suggesting that sometimes people do odd things when a robot asks them to, such as throwing out expensive textbooks and pouring orange juice on a potted plant. There are also a growing number of situations where humans put themselves at risk in the hands of a robot (e.g. autonomous cars, drones flying overhead, robot security guards, etc.). We thus wanted to develop a framework that allows robots to understand trust relationships with humans."

The conceptual framework developed by Wagner and his colleagues generated several testable hypotheses related to human-robot trust. In their study, the researchers examined these generated hypotheses and ran a series of experiments, gathering both evidence that supported their framework and conflicted with it.

"The trust framework started in Alan's research, which defined trust in a way that computers, and thus robots, can utilize," Robinette said.

In his previous work, Wagner defined situational 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 trustor has put its outcomes at risk'. His definition primarily focuses on the risk involved in a given situation, also highlighting the belief that a person/robot will act to reduce the risk for the other person/robot.

"Our framework provides criteria for what constitutes a trust situation and gives several categories of situations that do not require trust, for

instance where there is no risk to the trustor or where the risk cannot be mitigated for one reason or another," Robinette said. "With this framework, a robot can evaluate for itself whether the situation it is in requires trust or not, and then act appropriately."

The researchers tested the hypotheses generated by their framework in a series of tests and experiments. For instance, in one experiment, they presented a group of human participants with scenarios that involved trusting or not trusting someone else, then asked them which of these two options they would choose. The participants agreed with the conditions for trust generated by their framework, to a very high degree.

The researchers carried out several other experiments that evaluated the accuracy of the hypotheses generated by their conceptual framework. Some of these gathered evidence supporting these hypotheses, while a few produced conflicting results.

"I think the most meaningful finding from this work is that we found considerable supporting evidence for this framework in studies with many participants from diverse backgrounds," Robinette said. "This means that the trust framework can be used in most situations, allowing robots to better understand why humans around them are acting as they do. The robot may even be able to use this framework to steer humans to less risky situations, for instance by recognizing that a person is placing too much trust in a robot, perhaps to do something it was not programmed for, and informing the person of their error."

The framework devised by Wagner and his colleagues could be applied to a variety of situations that involve trust between humans and robots. Howver, in some cases the framework's hypotheses were not sufficiently accurate, for instance when people were asked to trust a robot in what appeared to be an emergency situation.

These findings are valuable nonetheless, as they shed light on specific areas in which people find it harder to trust robots. Future research could take a closer look at why participants made these choices and what prevented them from trusting robots, while also exploring ways in which robotics engineers can enhance safety around robots.

"Alan and Ayanna have been working to extend this research into the healthcare robotics domain," Robinette said. "I believe that Alan also has a project to investigate emergency evacuation robots and their relationship to people in greater detail. I have recently been working on human-machine teaming and plan to apply this trust framework to the relationship between humans and the robots that they work with."

More information: Modeling the human-robot trust phenomenon: a conceptual framework based on risk. DOI: 10.1145/3152890. dl.acm.org/citation.cfm?id=3152890