

The Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations

Paul Robinette^{1,2}, Alan R. Wagner², Ayanna M. Howard¹

¹Georgia Institute of Technology

²Georgia Tech Research Institute

Tech Report GT-IRIM-HumAns-2015-001

Acknowledgements

Partial support for this research was provided by Air Force Office of Sponsored Research contract FA9550-13-1-0169.

Partial support for this research was provided by the Motorola Foundation Professorship.

Corresponding author: Paul Robinette, probinette3@gatech.edu

Précis

We vary the ability of robots to mitigate a participant's risk in a navigation guidance task to determine the effect this has on the participant's trust in the robot in a second round. A significant loss of trust was found after a single robot failure.

Keywords

Trust, robot, human-robot interaction, social robots, rescue robots

ABSTRACT

Objective: This paper explores the impact that a robot's initial performance has on a person's decision to trust the robot later.

Background: As robots become more capable, the prospect that people will begin to trust these machines and place themselves at risk with respect to the machine increases. For robots tasked with guiding people in emergencies, failure on the part of the robot may result in the person's death. It is therefore important that we understand how a robot's performance influences the decisions a person makes and how different facets of the situation, such as perceived risk, play a role.

Method: We conducted human subject experiments that task individuals with navigating a maze within a fixed amount of time. The person had the option of using a robot for guidance in each of two separate navigation rounds. The robot performed poorly (with respect to time) in half of the cases and well in half of the cases. The participant's decision to use the robot and self-report of trust in the robot served as dependent measures.

Results: We found 53% of individuals self-report less trust when the robot performs poorly. In a simulated emergency situation 50% self-reported less trust after poor performance. Self-reports of trust were strongly correlated to the decision use the robot for guidance, $\phi(90) = +0.745$.

Conclusion: A robot's initial performance is vital for continued trust. Poor initial performance has a strong negative impact on a person's trust in the robot.

Application: This article should influence robot testing and design.

INTRODUCTION

Trust is a requirement in every interaction that involves risk, from everyday tasks to life-and-death situations. Victims in emergencies do not waste precious time arguing with firefighters and other emergency responders; they follow the responders' directions because they trust the agent providing guidance. In the same way, lost shoppers will follow the directions of a helpful passerby because the shoppers trust that the passerby has no interest in leading them astray. The risk involved in these two scenarios is different, but the concept of trust is inherent in both.

Robots have incredible potential to assist humans in everyday and emergency tasks. One such task is in aiding victims during a fire. Concerned about high casualty rates in emergency situations such as the Station Nightclub Fire of 2003, we have, in previous research, explored numerous situations where emergency guidance robots can improve human survivability in evacuations (Robinette & Howard, 2011; Robinette, Vela, & Howard, 2012; Robinette, Wagner, & Howard, 2013). We have also developed several robotic platforms that are capable of communicating understandable guidance instructions to humans (Robinette, Wagner, & Howard, 2014). These guidance robots can be deployed in a variety of scenarios outside of the emergency domain, including the lost shopper example discussed above.

Today, robots are being actively deployed in scenarios that help humans achieve tasks ranging from cleaning floors to bomb disposal; however such tasks either present low risk to humans (e.g. cleaning a floor) or are tightly controlled by

human experts (e.g. bomb disposal). To increase the potential for autonomous robots to aid humans in additional high-risk tasks, humans must first trust the robots to perform these tasks correctly. Exploring the conditions that result in a decision by humans to trust or not trust a robot is critical to allowing robots to become more useful to humans.

To develop trustworthy robots, we must first examine the conditions that affect a human's decision to trust a robot. One condition is prior task performance. In this paper, we ask: **how does the initial performance of the robot affect the human's decision to trust the robot later?** The understanding gained by exploring this question will allow researchers to create robots that humans are more likely to trust, develop robots that understand how to better manage a person's trust, and may provide insight into the phenomenon of trust itself. To answer this question, we have developed an interactive navigation simulation that allows participants to use a robot as a guide in a timed scenario. We measure the participant's decision to use the robot in an initial round, when the participant has little knowledge of the robot, and in a second round, after the participant has experience with the robot. In pilot studies, we found that participants generally accept robot guidance but are quick to reject the robot when it performs an unexpected action (Robinette & Howard, 2012) (Robinette, Wagner, & Howard, 2014).

Conceptualizing Trust

Numerous researchers have proposed conceptions of trust that range from computational implementations of cognitive processes (Castelfranch & Falcone, 2010), to neurological changes in reciprocity games (King-Casas, Tomlin, Anen,

Camerer, Quartz, & Montague, 2005), to a probability of an agent performing a particular action (Gambetta, 1990). Other researchers consider trust to have multiple forms, depending on the actors and environment (Hoffman, Johnson, Bradshaw, & Underbrink, 2013). After a review of the available literature, Lee and See conclude that trust is *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability* (Lee & See, 2004). Wagner builds on Lee and See's research with an operational definition of situational trust: trust is "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" (Wagner, 2009). In developing the navigation task, we sought to generate a situation in which people places themselves at risk with the belief that the robot will mitigate this risk. We felt that, if properly constructed, a scenario such as this could recreate an emergency situation where the person's survival depends on the actions of the robot.

Related research has focused on the factors that affect trust in a robot (Carlson, Desai, Drury, & Yanco, 2014). Carlson et al. finds that reliability and reputation impact trust in surveys of how people view robots. In contrast, we use simulations to record the person's actual behavior during an interaction involving trust. We focus on initial interactions with a robot, rather than trust that has been built over a long history.

Several measures of trust in a robot exist. Desai et al. asked participants to self-report changes in trust (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013).

Measurements of the frequency of operator intervention in an otherwise autonomous system have also been used (Gao, Clare, Macbeth, & Cummings, 2013).

Our study treats trust as a binary decision: either the person trusts the robot or the person does not. We wanted to investigate high-risk situations in which the person must make a decision to place their trust in a robot. We feel that this type of scenario better represents human-robot emergency situations. Moreover, many real and important situations present people with a decision that has little middle ground. For instance, being a passenger in an autonomous car is a binary decision in the sense that one chooses either to be a passenger or not be a passenger. The situation affords little opportunity for a third, middle risk option.

Hypotheses

In order to explore how a robot's initial performance affects a person's trust, we must measure the change in trust after the robot acts as a successful guide and after the robot does not act as a successful guide. Our first hypothesis examines this question directly:

H1. The self-reported trust will not be significantly lower in the second round if the robot performed well. Self-reported trust will be significantly lower in the second round if the robot did not perform well.

There are many ways for a robot to fail during a time-critical situation. For this guidance scenario, one failure mode is for the robot to be an inefficient or slow guide. This occurs when the robot successfully leads the person to the exit, but

requires a great deal of time to do so. Another type of failure is for the robot to not lead the person to the exit. One way to implement this type of failure is for the robot to stop moving somewhere within the maze. We hypothesize that:

H2. Participants that are guided by a robot that fails will self-report less trust than participants that are guided by a slow, inefficient robot.

As stated above, different measures for trust exist. One could use a measure of the person's behavior to infer the amount of trust. Alternatively, one could ask participants to self-report their trust. We hypothesize that:

H3. There will be a high correlation between participants who decide to use the robot in a round and participants who self-report that they trusted the robot.

Risk is a major component of trust (Wagner, 2009). Characteristics of the experimental scenario can influence a subject's perceived risk differently. For example, the risk associated by losing \$10 gambling will likely impact the behavior of people near poverty more than wealthy people. From an empirical point of view, we would like to control the factors that influence the subject's perceived risk. Yet, monetary incentives are a common method for putting a person at risk in order to explore trust (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005; Axelrod, 1984; Berg, Dickhaut, & McCabe, 1995). Our final hypothesis examines the

use of emergency scenarios as a possible replacement for monetary incentives in trust research.

H4. The decision to use the robot for guidance in the second round is significantly more sensitive to the robot's performance in the emergency scenario than in the bonus scenario.

METHODOLOGY

To address these hypotheses, two different experiments were conducted. Both experiments required a person to navigate a simulated maze with or without the help of a robot. In order to examine the impact that a robot's initial performance has on later decisions involving trust, the person was required to navigate a different maze in two separate rounds. They were given the option to use a guidance robot prior to navigating both mazes. Data reflecting their decision to use or not use a robot as well as surveys focused on the participant's reasoning was collected and used to confirm or refute the four hypotheses presented above.

As mentioned above, the decision to trust was a binary decision. The person either allowed the robot to provide guidance or did not. The decision to use the robot was viewed as an indicator of trust. We also measured trust by asking participants to self-report whether or not they agree with the statement: "I trusted the robot when I made my choice to follow or not follow the robot." In addition to the options to agree or disagree, we also gave an option labeled "Trust was not involved in my decision." In pilot studies, we found that some participants

interpret a disagreement to the statement to mean that they actively distrusted the robot, hence we provided a third option that clearly indicates they neither trust nor distrust the robot. The results focus on affirmations of trust.

Simulation Environment

One goal of this research is to better understand how people react in an emergency situation, so a simulation environment was created to resemble an office building. The simulation environment included corridors and rooms designed to give it a maze-like appearance (Figure 1). Participants were placed in the environment with no previous experience and required to find a single exit.

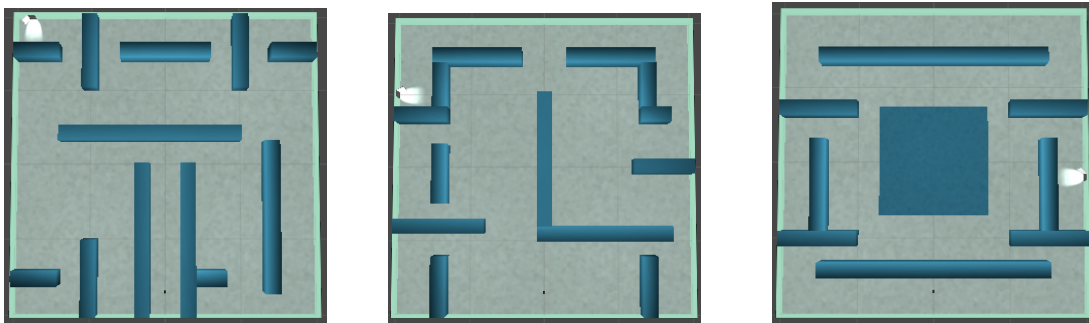


Figure 1: Overhead views of the three environments used in both experiments. Environments were designed to be similar to office layouts. Corridors and rooms were used to give maze-like qualities to make the simulation challenging.

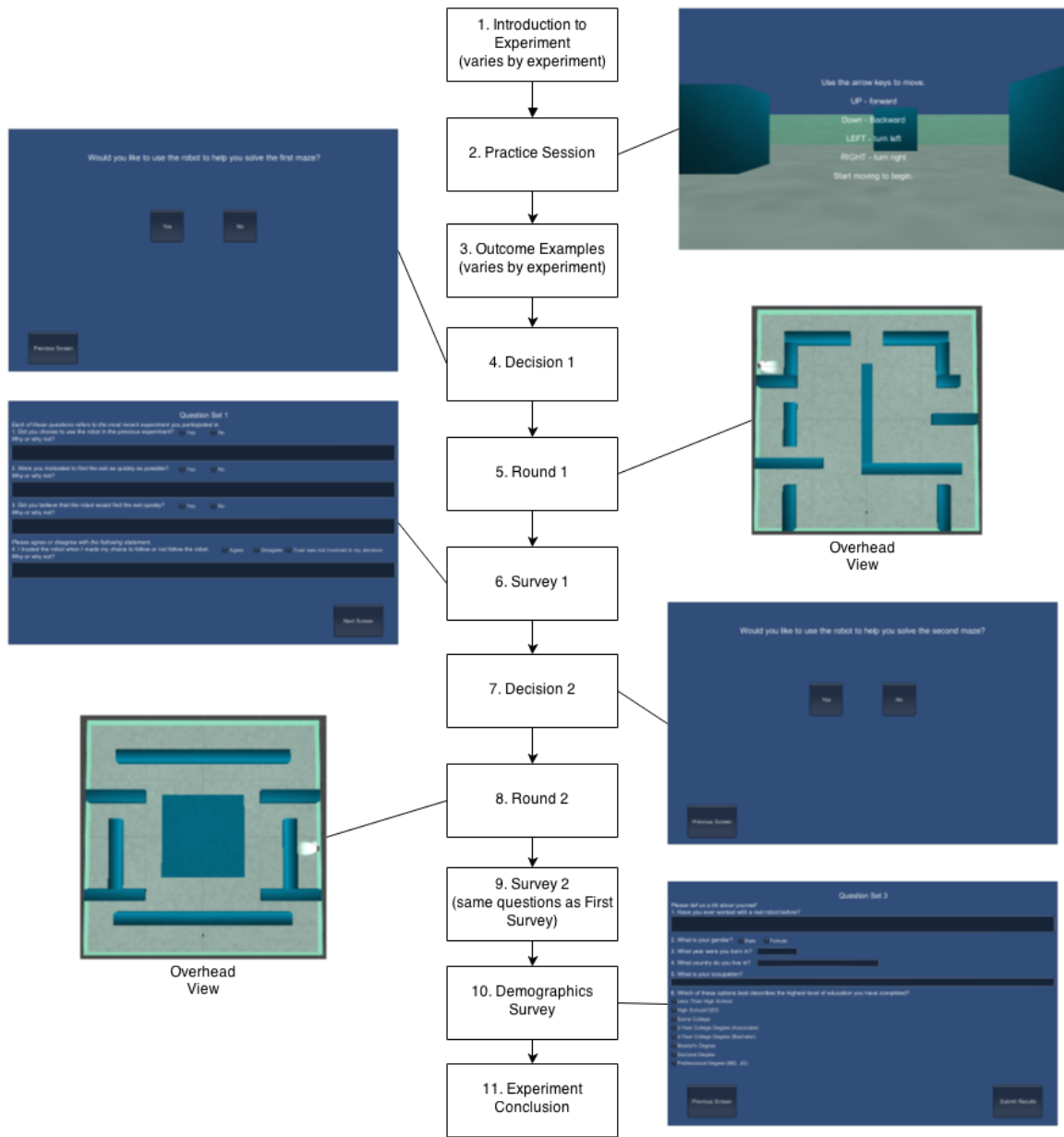


Figure 2: Experimental protocol with screenshots from experiment. The entire experiment was presented in a Unity 3D web game, including the survey questions.

The same general experimental setup was used for both experiments (Figure 2). Participants began each experiment by accepting the request on Mechanical Turk and clicking a link to a Unity 3D Web Player executable. Some participants had to download the Unity Web Player plugin to perform the experiment. Next they viewed an introductory message that described the navigation task they were to

perform. This page included photos of an exit and the guidance robot. The guidance robot varied in the two experiments. They were then offered the opportunity to practice navigating in a maze. They had a first-person view of the maze and used their keyboard arrow keys to move. After the practice session, they were presented with illustrative examples of prior human-robot performances in the maze. The nature of these examples varied with respect to the experiment. The participant was then asked to decide whether or not they would like a robot to provide guidance during the first round of the experiment. After making their choice the person then navigated the maze and completed a short survey (Figure 3). They were then offered another opportunity to decide if they wanted to use the guidance robot in the second round. They then navigated the maze in the second round and completed a short survey about their second round decision. The robot's guidance performance in the second round always matched its performance in the first round. The experiment concluded with a final survey that collected demographic information (Figure 4).

Question Set 1

Each of these questions refers to the most recent experiment you participated in.

1. Did you choose to use the robot in the previous experiment? Yes No
Why or why not?

2. Were you motivated to find the exit as quickly as possible? Yes No
Why or why not?

3. Did you believe that the robot would find the exit quickly? Yes No
Why or why not?

Please agree or disagree with the following statement.

4. I trusted the robot when I made my choice to follow or not follow the robot. Agree Disagree Trust was not involved in my decision
Why or why not?

Figure 3: Post-round survey presented to participants after each round. This survey gathered qualitative and quantitative information about a participant's trust in the robot as well as comments on their decision to use the robot or not.

Question Set 3

Please tell us a bit about yourself

1. Have you ever worked with a real robot before?

2. What is your gender? Male Female

3. What year were you born in?

4. What country do you live in?

5. What is your occupation?

6. Which of these options best describes the highest level of education you have completed?

- Less Than High School
- High School/GED
- Some College
- 2-Year College Degree (Associate)
- 4-Year College Degree (Bachelor)
- Master's Degree
- Doctoral Degree
- Professional Degree (MD, JD)

Figure 4: Demographics survey presented to participants after each experiment.

The quality of robot guidance (fast, slow, failed) was an independent variable studied in both experiments. The dependent variables for both experiments were 1) the participant's decision to follow or not follow the robot and 2) the participant's self-reported trust. Data on these variables was collected after each round for all participants.

Participant Inclusion and Exclusion Criteria

Crowdsourcing was used to collect data for both experiments. Crowdsourcing is a method for collecting data from a relatively large, diverse set of people (Paolacci, Chandler, & Ipeirotis, 2010). Crowdsourcing sites, like Amazon's Mechanical Turk, post potential jobs for crowdworkers, manage worker payment, and track worker reputation. The use of crowdworkers offers a quick and efficient complement to traditional laboratory experiments. Moreover, the population of

workers that provide the data tends to be somewhat more diverse than traditional American university undergraduates. In order to ensure the best possible data, participants were required to have a 95% acceptance rate for their previous work and were only allowed to participate once.

The experimental surveys required subjects to comment on the reasoning behind their decisions. Much of our previous work has indicated that participants understood our questions and thought logically about the answers (see (Robinette, Wagner, & Howard, 2014)). A participant's data was excluded if their comments were missing, nonsensical, or repeated throughout. Human participation in our experiments was approved by the Georgia Tech Institutional Review Board.

Robot Behavior

The actions of the robot inform the human of the robot's ability to be trusted in future interactions. H1 examines how the robot's behavior affects the participants' self-reports of trust in the second round. H2 explores different types of robot guidance failures: one that inefficiently leads the person to the exit and one that fails entirely to lead the person to the exit. In pilot studies we evaluated several different types of robot guidance failures. All but two of which were eliminated because participants were unable to determine that the robot had failed and hence resulting in an extremely long experiment completion time (see Table 1 for a listing of the robot guidance failure types that were not included in the experiments). Overall, three robot behaviors were defined that were used in the experiments:

- Fast navigation: the robot proceeds directly to the exit location (Figure 5). Robots that acted in this manner are capable of finding the exit within thirty seconds.
- Slow navigation: the robot explores many possible routes before eventually finding the exit (Figure 6). Robots that acted in this manner are capable of finding the exit in ninety seconds.
- Failed navigation: the robot proceeds directly to a corner of the environment that is not the exit location and then stops. This is meant to emulate the behavior of a robot that has incorrect information about the exit location. Robots that acted in this manner stopped moving after approximately thirty seconds at a point at least thirty seconds from the exit.

Table 1: Poor robot guidance behaviors that were explored as part of a pilot experiment.

Name	Description	Reason for Exclusion
Small Loops	Robot circled an obstacle continuously	Several loops around the obstacle were required before participants realized the robot had failed. The total time for the experiment was too long.
Large Loops	Robot circled a large area of the environment continuously	Participants could not realize that the robot had failed until it completed at least one loop. This could take several minutes by itself and thus the total time for the experiment was too long.
Continuous Searching	Robot searched through entire environment except location of actual goal position. After completing a search it started again.	Participants followed the robot for considerable time before realizing the robot had failed. Some participants would follow the robot for more than 15 minutes.
Wall Collision	Robot nearly found goal but then continuously collided with wall and was unable to proceed.	Participants did not understand that the robot was colliding with the wall and thus did not understand that it failed.

EXPERIMENT 1: BONUS MOTIVATION

Our first experiment used all three robot behaviors and both trust metrics to test the first three hypotheses. Successful and unsuccessful robots were used to test our first hypothesis (that participants will continue to trust a robot that performs well more often than a robot that performs poorly). The two unsuccessful robot behaviors (slow navigation and failed navigation) were included to determine if there was a difference in response (Hypothesis 2). Both of these hypotheses were tested as a between subjects experiment with the robot behavior as the independent variable so no participant saw more than one robot behavior type. Our dependent variable was the trust a participant had in the robot. We used two different metrics to determine this trust: their decision to put their outcomes at the risk of the robot's behavior and their self-reported trust in post-round surveys. The correlation between these two metrics serves as evidence for Hypothesis 3.

In keeping with the previous literature on trust research (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005), we used monetary risk to motivate participants' trust decisions. For this experiment, subjects were offered a \$1 bonus if they could find the exit of a maze within 30 seconds. After the first 30 seconds had elapsed the bonus began to decrease. Ninety seconds after the start of the experiment the bonus was \$0. Participants were informed that their choice to use a guidance robot or not would not directly affect their bonus in any way.

Experimental Setup

As noted above as well as in Figure 7 and Figure 8, the first stages of the experiment, the Introduction and Example Outcomes sections, were unique for each experiment. Each reflected the specific, monetary motivation in this experiment. We included one additional survey in this experiment to help us understand the motivations of participants on Mechanical Turk. This survey can be seen in Figure 9. The questions are unrelated to the hypotheses or research question and the survey was only included to help us design better experiments.

The first screen seen by the participants gave instructions for the simulation. The simulated environments were specifically referred to as “mazes” to give the participant an idea of their complexity and goal. The robot displayed during the introduction and used in the rounds was a Willow Garage TurtleBot 2. The 3D model of the robot was created out of CAD files distributed by the manufacturer. The practice session proceeded as described in the Methodology section.

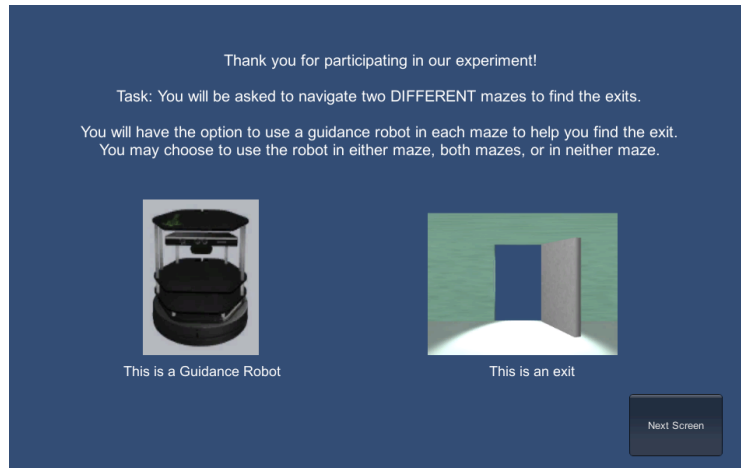


Figure 7: Introduction screen for the bonus motivation experiment. Images of the robot and the exit were used so that participants could gain familiarity with the simulation.

After the practice session, the participants were informed of the performance-based bonus and how to obtain it. Participants were given three example performances for the navigation task:

Example 1: The text “People who used a robot that quickly found the exit typically earned a bonus of about \$1.00” accompanied by a top-down view of a direct path to the exit in an example maze.

Example 2: The text “People who used a robot that did not quickly find the exit typically earned a bonus of about \$0.00” accompanied with a top-down view of a very indirect path to the exit in the same example maze.

Example 3: The text “People who did not use a robot typically earned a bonus of about \$0.50” accompanied with a top-down view of an indirect path to the exit in the example maze.

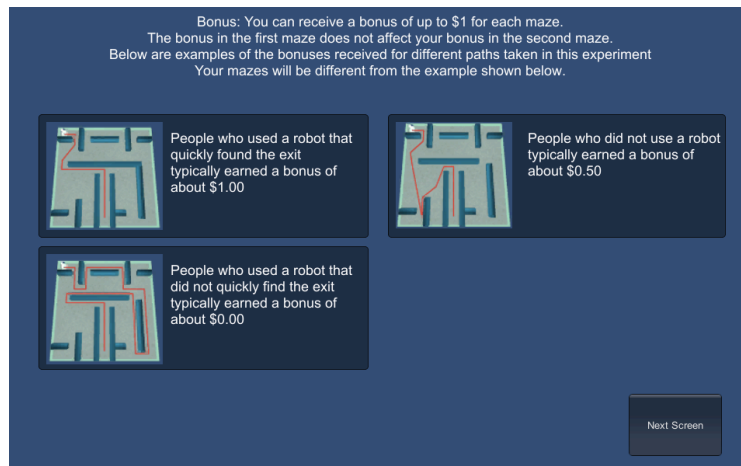


Figure 8: Example outcomes for the bonus motivation experiment. Participants were shown overhead views of an example environment with a path drawn in each so that they could understand the complexity of the environments as well as observe a visual representation of successful and unsuccessful robot behavior.

Question Set 1

Please rate your agreement with each of the statements below.

1. I would like to complete this experiment as quickly as possible, regardless of the amount of bonus I earn or how much fun I have.

Disagree - 1 2 3 4 5 6 Agree - 7

Please explain your answer below

2. I would like to earn the highest bonus possible in this experiment, regardless of the time it takes or how much fun I have.

Disagree - 1 2 3 4 5 6 Agree - 7

Please explain your answer below

3. I would like to have fun in this experiment regardless of the time it takes or the bonus I receive.

Disagree - 1 2 3 4 5 6 Agree - 7

Please explain your answer below

4. Please rate the following factors based on their importance to you in this experiment. 1 is most important and 3 is least important

The time it takes to complete the experiment 1 2 3

The bonus received at the end 1 2 3

The amount of enjoyment provided by this experiment 1 2 3

Previous Screen

Next Screen

Figure 9: Supplemental questions to determine motivations of participants. These helped in understanding the motivations of the participants and aided in the creation of our second experiment.

The participants were then asked whether they would like to use a robot in the first round or not. At the start of each round, the participants were reminded of the controls for moving in the environment and informed that their bonus was currently set at \$1.00 (Figure 10). When the participant began moving a timer in the top left of the screen started increasing, displaying the time spent navigating to a tenth of a second precision. The bonus was prominently displayed in the top right corner. After thirty seconds of navigating the maze, the bonus began to decrease at a rate of \$0.0167 per second (Figure 11). This meant that the bonus was completely depleted after ninety seconds. The second round was setup the same as the first but with a different maze. All other aspects of this experiment proceeded as described in the Methodology section.

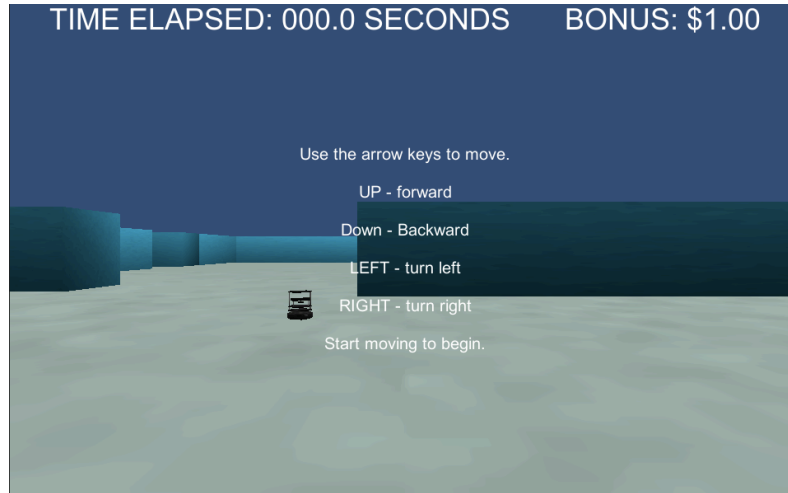


Figure 10: The image above presents the beginning of the first round of the bonus motivation experiment. Round 2 had the same text but took place in a different environment. Participants are reminded of the controls for the simulation, updated on the current amount of their bonus, and shown the time elapsed since the round began.

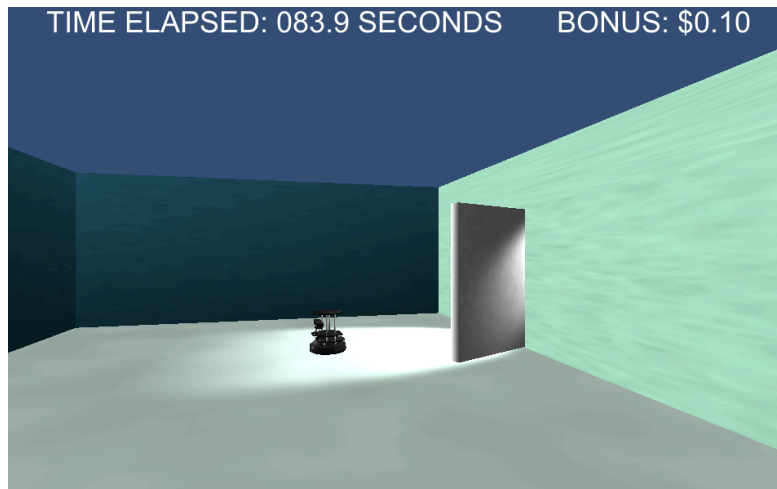


Figure 11: The figure presents a successful ending to round 1, albeit with a small bonus. This was typical of subjects being guided by a slow robot.

Because participants had no control over the amount of bonus they earned; they were all paid the full \$2.00 bonus after their experiment was completed. This information was not made available to any participant before they performed the experiment.

Results

A total of 106 participants (mean age=31.0, 60.4% male) completed the first experiment, 84.9% of which chose to follow the robot in the first round, with no prior knowledge of the robot's behavior. Figure 12 depicts the number of participants who used the robot in rounds 1 and 2 for the fast and slow/failed robot behaviors and the self-reported trust in rounds 1 and 2 for the different robot behaviors. Only participants who chose to follow the robot in round 1 are reported. As can be seen in the figure, self-reported trust decreases significantly (53%, SE= 6.3%) when the participants experience a slow or failed robot in the first round, $\chi^2(1, 90) = 12.86, p < 0.001$. Only a 4% (SE = 7.3%) decrease in trust was reported by participants that were guided by a fast robot. The decision to follow the robot, however, does not decrease significantly, $\chi^2(1, 90) = 1.87, p = 0.172$. Figure 13 shows the results for the different failure modes. The type of robot failure had no impact on either the self-reported trust (0% difference) or the decision to follow (0% difference). In both the first and second round a strong positive correlation was found between following the robot and reporting trust in the robot, $\phi(106) = +0.628$ for round 1 and, $\phi(90) = +0.422$ for round 2.

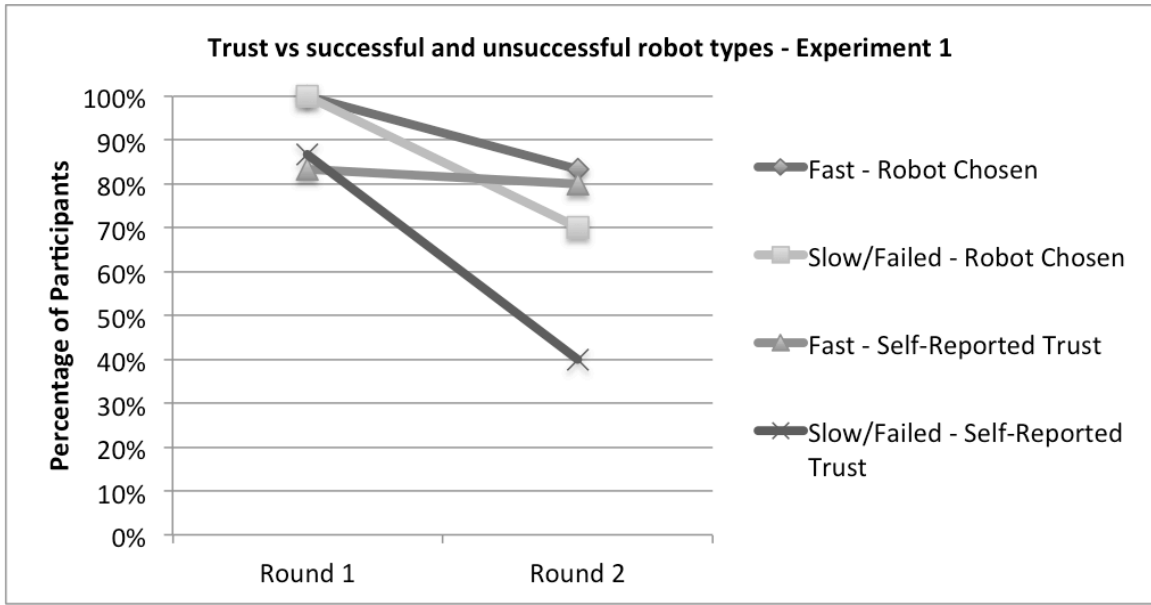


Figure 12: Change in decision to use robot and self-reported trust between the two rounds for the successful and unsuccessful robots. Note that a majority of participants continued to use the slow/failed robots even though half had lost their trust in the robot.

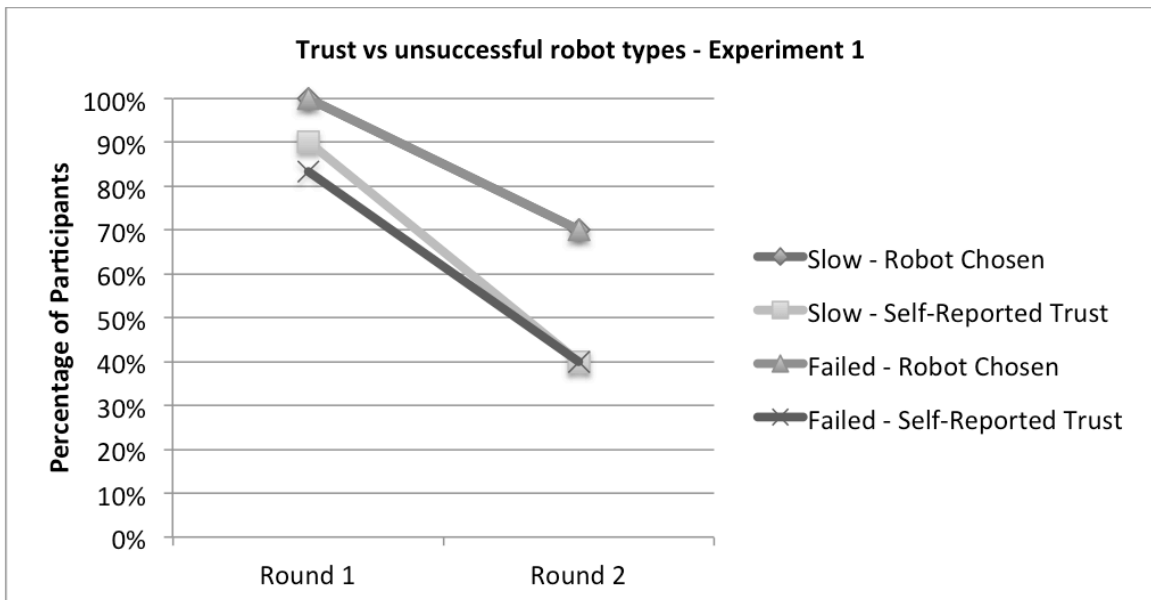


Figure 13: Change in decision to use robot and self-reported trust between the two rounds for the slow and failed robots. The same number of participants chose to use each and the same number reported trust in each.

Discussion

Overall, the results strongly support some of our hypotheses and do not support others. With respect to H1 the data indicates a 53% decrease in self-reported trust when the robot fails versus a 4% decrease when the robot does not fail. This result supports our hypothesis that self-reported trust significantly decreases after the robot provides slow or failed guidance. This result is important in that it shows that only a single failure can strongly and quickly influence a person's trust in the robot, which may have ramifications on the testing and evaluation of such systems. It is also noteworthy that the majority of people (84.9%) chose to follow the robot initially. This result appears to imply that people tend to trust initially.

Our second hypothesis focused on the manner in which the robot failed. We predicted that a robot that fails by traveling a short distance and stopping would have a significantly larger negative impact on both self-reported trust and the subsequent decision to follow than a robot that merely slowly led to the exit. The data shows that this is not the case. Our results indicate that either failure impacted trust equally both with respect to the self-report and subsequent following. The fact that there was 0% difference in both cases is presumably a statistical artifact.

The result is intriguing. Does it indicate that the reduction in trust is only a reflection of good versus bad performance and not, as we expected, mitigated by the way in which a robot failed? This could be an area for significant further research.

Our third hypothesis examined the relationship between self-reported trust and the decision to follow the robot. We predicted that both the participant's trust

and the likelihood of following would be strongly correlated. Indeed, we found a strong positive correlation between following the robot and self-reporting trust. Yet, the results show that numerous participants (26% of all participants) choose to follow the robot in the second round even though they reported not trusting it.

We examined the survey comments to better understand each participant’s rationale. Table 2 summarizes the most common comments from round 2. Note that, of the people that were guided by a slow or failing robot, many choose to follow the robot in the second round because they believed that the robot’s help was better than no help at all (n=7) or they thought that the robot would perform better this time (n=5). These comments hint that participants were deciding to follow the robot in spite of the loss of bonus.

Table 2: Summary of comments from Experiment 1.

Robot Behavior	Used Robot?	Self-reported trust	Comment Description
Fast (n=30)	Yes (n=25)	Positive (n=22)	Robot performed well (n=21)
			Did not trust robot, trusted programmers (n=1)
		Negative/Neutral (n=3)	Impossible to trust machine (n=1)
			Trusted robot initially but explored on own instead of completing maze (n=1)
	More than two examples required to trust something (n=1)		
	No (n=5)	Positive (n=2)	No complaint about robot, wanted to try experiment for themselves (n=2)
			No complaint about robot, wanted to try experiment for themselves (n=1)
Negative/Neutral (n=3)		Thought robot would perform worse in second round (n=1)	
Slow (n=30)	Yes (n=21)	Positive (n=11)	Robot performed better than human alone (n=7)
			Did not realize robot performed poorly (n=3)
			Thought robot would perform better in second round (n=1)
		Negative/Neutral (n=10)	Curiosity (n=6)
			Robot performed better than human alone (n=1)

	No (n=9)	Positive (n=1)	No complaint about robot, wanted to try experiment for themselves (n=1)
		Negative/Neutral (n=8)	Robot performed poorly (n=7)
	Yes (n=21)		Positive (n=11)
		Negative/Neutral (n=10)	
	No (n=9)		Positive (n=1)
		Negative/Neutral (n=8)	
			<i>Unclear response</i> (n=1)
			Robot performed poorly (n=8)

We performed an analysis on our motivational survey to better understand the participants. About half of participants (55) reported that their most important motivation with respect to the experiment was money. The rest were evenly divided between time (25) and fun (24). These results indicate that participants are not solely motivated by simple monetary bonuses in the experiment. Hence, some choose to follow the robot in the second round in spite of its failure and the fact that they self-reported not trusting it because they believed it would ultimately be faster or more fun to follow the robot.

Experiment 1 Conclusion and Motivation for Experiment 2

The experiment supports our hypothesis that self-reported trust decreases after poor guidance. Further, we found a strong correlation between self-reported trust and the subject's decision to follow the robot. No statistically significant difference resulted from the different types of robot guidance failure. Still, the data indicates that participants were significantly more likely to follow the robot in spite of a failure and a lack of trust. An examination of the participant's comments hinted

that the participants are motivated by factors other than simply the bonus, such as time and fun.

These survey comments and our study of the participants' motivations led us to develop a second experiment that sought to better align the participants' motivations with the task goals. This second experiment asked participants to act as if they were in an emergency. Instead of receiving a bonus, a quick exit from the building rewarded them with "survival." We hypothesized that being motivated by an emergency would result in a better alignment between their self-reported trust and their decision to follow the robot.

EXPERIMENT 2: EMERGENCY MOTIVATION

For this experiment, participants were told that our goal was to discover how people leave a building in an emergency. Instead of receiving a bonus for a fast completion, they were told that they would only survive if they found the exit in time. During the rounds, a countdown timer appeared in the middle of their view to tell them the remaining time. As with the previous experiment, this study was conducted using the Unity simulation and Amazon's Mechanical Turk. Participants were compensated \$2.00 for their participation in this experiment.

Experimental Setup

There were several differences between this experiment and Experiment 1. First, participants were told in the introduction screen that "We are testing how people leave buildings in emergencies" and asked to

“Please act as you would in a real emergency!” (Figure 14). The word “building” was used instead of “maze” to further reinforce the emergency portion of the simulation.

The robot in this experiment was a TurtleBot 2 modified with two PhantomX Pincher AX-12 arms to allow it to gesture. The robot was also given signage to indicate that it is an emergency evacuation robot. The arms waved while it moved to attract attention. The robot was also given signage to indicate that it is an emergency evacuation robot. The arms waved while it moved to attract attention. The robot’s appearance and gestures were evaluated in a previous paper and it was found that participants understood it better than other forms of evacuation robots (Robinette, Wagner, & Howard, 2014).

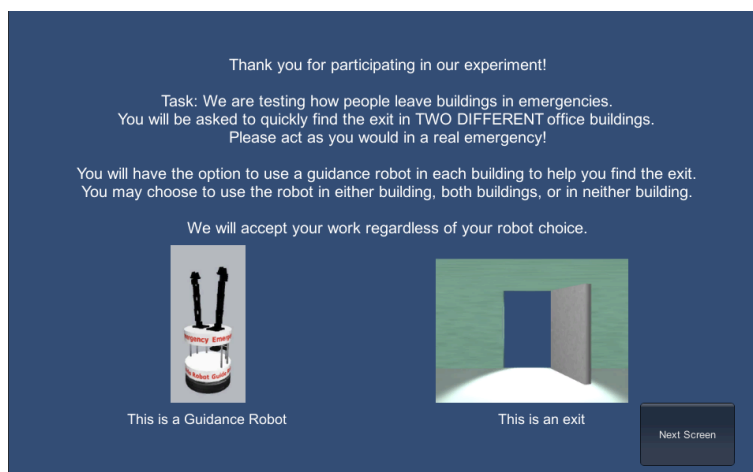


Figure 14: Introduction to emergency motivation experiment. Note that the robot is different from in Experiment 1. Additionally, participants were told that this experiment was to determine how humans evacuate buildings, not how humans interact with robots.

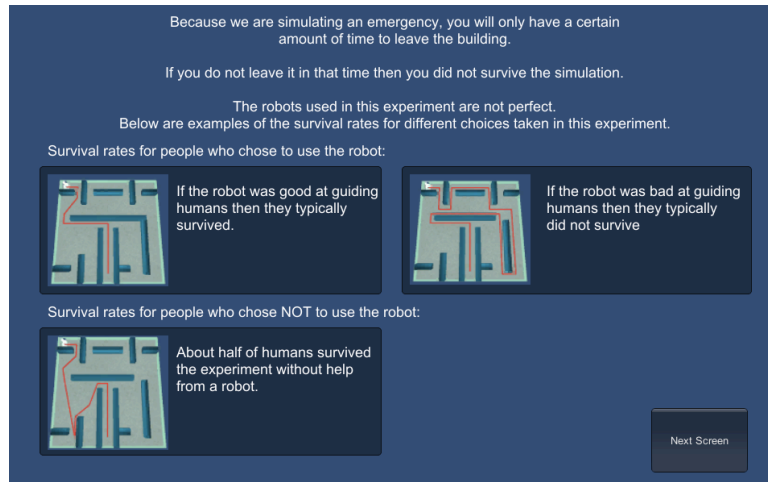


Figure 15: Example outcomes in emergency motivation experiment. Again, participants were shown overhead views of the example environment with a variety of paths, but this time they were presented with survival possibilities, not monetary rewards.

For this experiment each round ended after 60 seconds regardless of the participant's ability to find the exit. Once again, before selecting whether or not to use the robot, the participant was presented with a series of example experimental outcomes. The examples reflected the change to an emergency scenario (See Figure 15). The examples were:

Example 1: The text "If the robot was good at guiding humans then they typically survived." accompanied by a top-down view of a direct path to the exit in an example maze.

Example 2: The text "If the robot was bad at guiding humans then they typically did not survive" accompanied with a top-down view of a very indirect path to the exit in the same example maze.

Example 3: The text “About half of humans survived the experiment without help from a robot.” accompanied with a top-down view of an indirect path to the exit in the example maze.

The words “EMERGENCY! Please leave the building now! EMERGENCY!” as well as the time remaining to exit (to a tenth of a second precision) appeared in the top-center of the participants’ view throughout the entire round. See Figure 16 for an example of the beginning, Figure 17 for an example of a successful ending and Figure 18 for an example of an unsuccessful ending to the experiment.

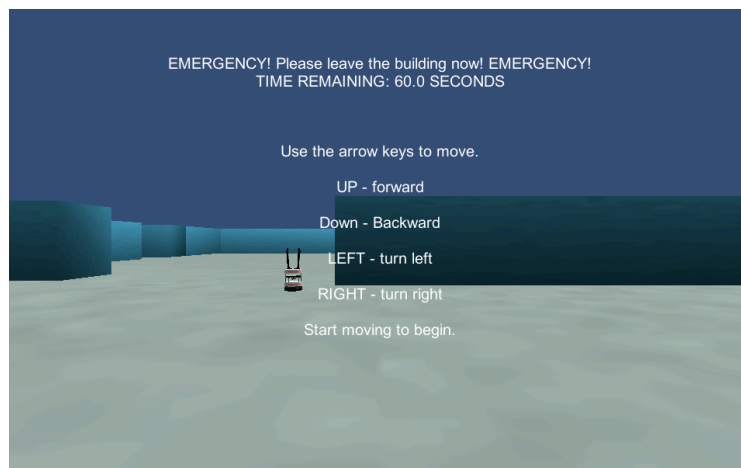


Figure 16: Beginning of the first round in the emergency motivation experiment. The timer counted down instead of up in this experiment and was moved to the center of the screen for maximum visibility. Text to indicate that this is an emergency was also placed on the screen.



Figure 17: An example of a successful exit in the emergency motivation experiment. This typically happened only in the fast robot conditions.



Figure 18: An example of an unsuccessful exit in the emergency motivation experiment. Text informed the participant there was no time remaining. The robot can be seen in the distance.

Other than these changes, the experiment was identical to the first experiment. Participants were again required to complete the same survey examining their trust in the robot and reasoning for choosing the robot. This experiment also consisted of two rounds.

Results

A total of 129 participants (mean age=31.8, 60.5% male) completed the second experiment, 69.8% of which decided to use the robot in the first round. As shown in Figure 19, the decision to follow the robot decreases by 50% (SE = 6.5%) in the second round when the participant interacts with a slow/failed robot in the first round. Moreover, the percentage of people that follow is only 3% (SE = 3.3%) less in the second round when the person encountered a fast robot in the first round. Self-reported trust follows a similar trend with trust decreasing 53% (SE = 6.2%) when participants experienced a slow/failed robot and self-reported trust increasing by 3% (SE = 5.5%). These results are statistically significant for both metrics, for the following decision, $\chi^2(1, 90) = 19.29, p < 0.001$, and for the trust self-report, $\chi^2(1, 90) = 24.31, p < 0.001$. Figure 20 shows the results for the different failure modes. The type of failure had minimal impact in the participant's decision to follow (47%, SE = 9.1% chose to follow the slow robot, 53%, SE = 6.3% for the failed robot, $\chi^2(1, 60) = 0.27, p = 0.606$). There was also a minimal change in self-reported trust (slow robot: 40%, SE = 9.1%, failed: 30%, SE = 5.9%, $\chi^2(1, 60) = 1.15, p < 0.284$). A strong positive correlation was found between choosing to use the robot and reporting trust in the robot in both rounds: $\phi(129) = +0.661$ for round 1 and $\phi(90) = +0.745$ for round 2.

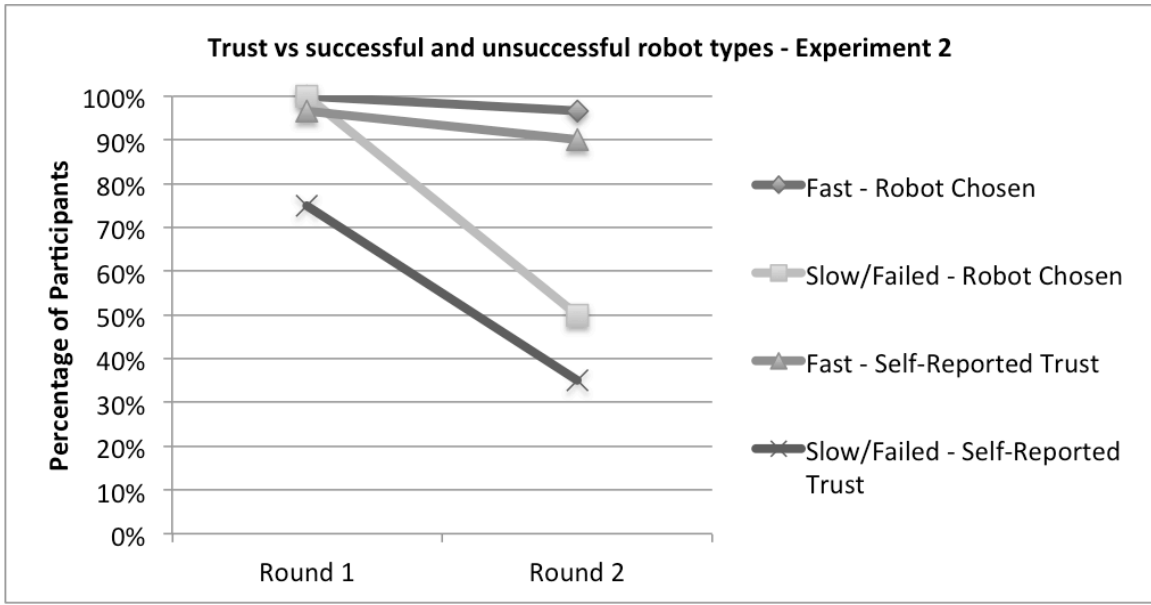


Figure 19: Change in decision to use robot and self-reported trust between the two rounds for fast and slow/failed robots. Note that the decision to use the robot dropped with self-reported trust in this experiment, unlike in Experiment 1.

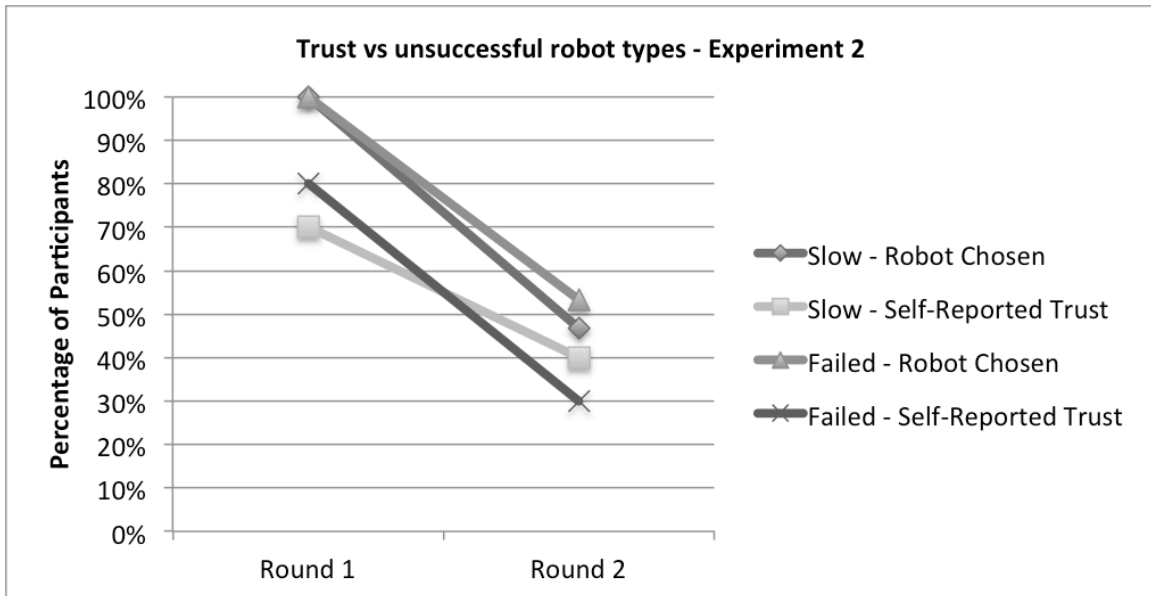


Figure 20: Change in decision to use robot and self-reported trust between the two rounds for the slow and failed robots. While the results are not identical in this round, as they were in Experiment 1, they are still not statistically significant.

Discussion

The results from this experiment strongly support H1, H3 and H4 but do not support H2. A single failure of a robot caused 50% of participants to stop using the robot in the second round, compared to just a 3% drop with a successful robot. This supports our hypothesis that participants will continue to trust a robot that performs well (H1).

Our data indicates that a smaller percentage of participants chose to use the robot in the first round when compared to the first round of Experiment 1. While a majority still chose to use the robot, and thus our findings from previous work are still supported, we did not expect such a change. Many participants justified their choice by stating that they did not want to put their life in the hands of a machine. This indicates that people are more likely to initially trust a robot when there is a lower risk (e.g. a financial risk instead of a survival risk). This data serves as evidence that people take the emergency scenario, and the risk it entails, seriously.

With respect to the type of robot failure, both experiments showed no difference in either self-reported trust or the decision to use the robot if the person experienced a slow robot versus a robot that stopped moving before arriving at the exit. This is an interesting area for future work as it indicates that participants do not discriminate based on how the robot failed, only that it did fail.

The results from this experiment show an even greater correlation between self-reported trust and the decision to use the robot than was seen in Experiment 1. This supports our third hypothesis: self-reported trust and the decision to use the

robot are correlated. Only 12% of participants chose to follow a robot that they did not report trusting in the second round of this experiment.

We also found strong support for our fourth hypothesis: the decision to use the guidance information from the robot was more sensitive to the behavior of the robot in the emergency scenario than in the bonus scenario. This result suggests that an emergency scenario, in contrast to a bonus scenario, does influence participants to act in a manner that is aligned with their self-reported trust.

Again, motivations for participants' actions and reports can be found in the comments. A short description of a selection of these comments can be found in Table 3. Note that not all participants' comments are included in this table for brevity and some participants gave multiple reasons for their actions.

Table 3: Summary of comments from Experiment 2.

Robot Behavior	Follow Decision	Trust Answer	Comment Description
Fast (n=30)	Yes (n=29)	Positive (n=27)	Robot performed well (n=24)
		Negative/Neutral (n=2)	Logical choice, not trust (n=1)
			Decided to proceed on own for fun after choosing to use robot (n=1)
	No (n=1)	Positive (n=0)	
Negative/Neutral (n=1)		Thought robot would perform worse in second round (n=1)	
Slow (n=30)	Yes (n=15)	Positive (n=12)	Curiosity (n=5)
			Thought robot would perform better in second round (n=3)
			Robot moved quickly, and thus was trustworthy (n=2)
			Did not realize robot performed poorly (n=2)
	Negative/Neutral (n=3)	Curiosity (n=3)	
		No (n=15)	Positive (n=1)
Negative/Neutral (n=14)	Robot performed poorly (n=13)		
	No complaint about robot, wanted to try		

			experiment for themselves (n=2)
Failed (n=30)	Yes (n=16)	Positive (n=9)	Robot performed better than human alone (n=6)
			Thought robot would perform better in second round (n=3)
		Negative/Neutral (n=7)	Curiosity (n=5)
	Robot performed better than human alone (n=2)		
	No (n=14)	Positive (n=0)	
			Robot performed poorly (n=12)
Negative/Neutral (n=14)		No complaint about robot, wanted to try experiment for themselves (n=2)	

The comments indicate that participants took the emergency scenario seriously. Several comments note that individuals acted as if they felt real pressure to find the exit quickly (one participant wrote “It felt like a challenge, and I treated it as an emergency as instructed,” another wrote, “Burning building, needed to get out”). Some likened it to getting the high score in a video game while others just wanted to “survive” the simulation. Participants who did not successfully survive the first round typically stated that they were upset with the outcome. Some were upset at their robot, some at themselves. Almost all participants who failed to survive in the first round vowed to live in the second. We believe these comments are evidence that using simulated emergency scenarios fosters a sense of risk in the participant that is critical for human-robot trust experiments.

Experiment 2 Conclusion

Overall this experiment has broadened support for several of our central hypotheses and demonstrated the validity of a new scenario for investigating human-robot trust. The importance of having viable scenarios to examine human-robot trust should not be overlooked. Most researchers agree that risk is a central component of trust (Rousseau, Sitkin, Burt, & Camerer, 1998). Yet, developing

scenarios that place participants under a type of risk that is simultaneously acceptable to internal review boards and able to be mitigated by a robot is challenging. Typically, investment games with a bonus risk have been used (King-Casas, Tomlin, Anen, Camerer, Quartz, & Montague, 2005; Axelrod, 1984; Berg, Dickhaut, & McCabe, 1995). The results presented here indicate that an emergency evacuation scenario might also be a viable method to explore trust.

PRACTICAL IMPLICATIONS

The practical implications of this work include many suggestions for future robot development. In these experiments, we found that a majority of people were willing to accept guidance information from a completely unknown robot, even when placed in a simulated emergency situation. We also found that the participants' trust would immediately change once the robot fails to meet their expectations. From a methodological stand point, this work could influence the way trust in robots is investigated. Our results show that simulated emergency situations result in participant behavior which closely matches their self-report. The results also show that this is not so for some types of financial incentives. Overall, this work begins to explain how and why a robot's behavior impacts a person's trust. Hence, we believe that our findings will have broad implications for a variety of robot applications.

OVERALL CONCLUSION AND FUTURE WORK

This article has explored how a person's trust in a robot is affected by failure. Our results show that people will often initially trust an unknown robot. Still, even a single failure strongly impacts a person's trust. Furthermore, we found that the

manner in which the robot fails does not matter. Yet, in some cases people may still act as if they trust the robot even if they self-report that they have little trust. The reason for this seems to lie with the experimental scenario. Experiments which attempt to equate the person's risk to a bonus appear to underestimate other motivations such as time and fun. On the other hand, a simulated emergency scenario does result in a close correspondence between a participant's actions and self-reported trust.

This work is not without limitations. The results are likely influenced by the fact that users conducted the experiment using an internet-based simulation. Subjects were presumably in a relatively safe locale, such as their home. Hence, an experiment which asks subjects to pretend that they are in an emergency is unlikely to generate the same adrenaline and emotional state as an actual emergency. Although many participants reported a strong desire to achieve the maximum bonus in Experiment 1 and to "survive" in Experiment 2, we cannot be sure that their decision making in these simulated environments matches what their behavior would be in a real emergency. Laboratory experiments that recreate an emergency might address some of these limitations.

There is still considerable future work to be done. Some participants continued to use a poorly performing robot in spite of obvious failures. It may be valuable to explore how long individuals will continue to trust a failing robot. Additionally, developing methods that allow a robot to communicate when it should or should not be trusted appears valuable. This paper specifically explores trust decisions at discrete points; however, trust in these situations tends to be a

continuous decision. Experiments which track a person's trust in the system continuously as it fails and attempts to recover could provide additional insights.

As robots enter everyday life, we must be conscious of their affect on the humans they are supposed to help. Currently, robots aid humans by performing low-risk tasks, such as cleaning floors, but in the future they may be capable of saving human lives in high-risk situations, such as emergency evacuations. This paper gives insight into the situational factors and robot behaviors that impact a human's decision to trust a robot. These results can aid in the creation of trustworthy robots and provide data that can eventually be used to teach robots how to increase or decrease their trustworthiness dynamically.

KEY POINTS

- Robots were seen as trustworthy guides in both a monetary reward motivated scenario and a simulated emergency scenario until their first failure.
- The mode of failure did not affect the number of participants who stopped trusting the robot.
- Participants reported that they acted as if the simulated emergency was a real emergency and this is corroborated by a large drop in their decision to continue using a failed robot in the second round of the experiment.
- Self-reported trust in the robot and the participant's decision to use the robot had a strong positive correlation in both rounds of both experiments. It was strongest in the second round of the simulated emergency experiment, indicating that this scenario caused participant's motivations to best align with the risk of the scenario.

REFERENCES

- Axelrod, R. (1984). *The Evolution of Cooperation*. . New York: Basic Books.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games Economic Behavior* , 10, 122-142.
- Carlson, M. S., Desai, M., Drury, J. L., & Yanco, H. A. (2014). Identifying Factors that Influence Trust in Automated Cars and Medical Diagnosis Systems. *Proceedings of the AAAI Spring Symposium on the Intersection of Robust Intelligence and Trust in Autonomous Systems*. Palo Alto, CA: AAAI.
- Castelfranch, C., & Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. New York, NY: Wiley Publishers.
- Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 251-258). Tokyo, Japan.
- Gambetta, D. (1990). Can We Trust Trust? In D. Gambetta (Ed.), *Trust, Making and Breaking Cooperative Relationships* (pp. 213-237). Oxford England: Basil Blackwell.
- Gao, F., Clare, A. S., Macbeth, J. C., & Cummings, M. L. (2013). Modeling the Impact of Operator Trust on Performance in Multiple Robot Control. *AAAI Spring Symposium: Trust and Autonomous Systems*. Palo Alto, CA: AAAI.
- Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in Automation. *Intelligent Systems* , 28 (1), 84-88.

- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in Two-Person Economic Exchange. *Science* (308), 78-83.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, pp. 50-80.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5 (5), 411-419.
- Robinette, P., & Howard, A. M. (2011). Incorporating a model of human panic behavior for robotic-based emergency evacuation. *RO-MAN, 2011 IEEE* (pp. 47-52). IEEE Press.
- Robinette, P., & Howard, A. M. (2012). Trust in emergency evacuation robots. *Safety, Security, and Rescue Robotics (SSRR), 2012 IEEE International Symposium on* (pp. 1-6). IEEE.
- Robinette, P., Vela, P. A., & Howard, A. M. (2012). Information propagation applied to robot-assisted evacuation. *Robotics and Automation (ICRA), 2012 IEEE International Conference on* (pp. 856-861). IEEE.
- Robinette, P., Wagner, A. R., & Howard, A. (2013). Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. *AAAI Spring Symposium, Stanford University*, (pp. 78-83). Palo Alto.
- Robinette, P., Wagner, A. R., & Howard, A. M. (2014). Assessment of Robot Guidance Modalities Conveying Instructions to Humans in Emergency Situations. *RO-MAN, 2014 IEEE*. IEEE.

Robinette, P., Wagner, A. R., & Howard, A. (2014). Modeling Human-Robot Trust in Emergencies . *AAAI Spring Symposium, Stanford University*.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so Different After All: A Cross-Discipline View of Trust. *Academy of Management Review* , 23, 393-404.

Wagner, A. R. (2009). *The Role of Trust and Relationships in Human-Robot Social Interaction*. Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.

Paul Robinette is PhD student in the Robotics program at Georgia Institute of Technology. He received a M.S. in Computer Engineering in 2010, a B.S. in Physics in 2008 as well as B.S. in Computer Engineering in 2010 from the Missouri University of Science and Technology. He has worked with the Georgia Tech Research Institute for the last three years on projects ranging from autonomous underwater vehicles, to micro ground vehicles, to human-robot trust. His primary research is on the use of robots to guide and advise victims in emergency situations.

Alan Wagner's research interests include social robotics, human-robot interaction, computational social systems, game theory, trust and deception. He received a B.A. in Psychology from Northwestern University in 1996, a M.S. in Computer Science from Boston University in 2001 and a Ph.D. in Computer Science from Georgia Institute of Technology in 2009. His work focuses on the development of a framework for social action selection for robots based on theories from Social Psychology. His research on deception was described as the 13th most important invention of 2010 by Time Magazine. Other awards include the best paper award at RO-MAN 2007, the Air Force Young Investigator Award (2013), and Georgia Research Institute's Innovative Research Award (2013). He holds the position of senior research scientist at Georgia Institute of Technology's Research Institute.

Dr. Ayanna Howard is the Motorola Foundation Professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. She received her B.S. from Brown University, her M.S.E.E. from the University of Southern California,

and her Ph.D. in Electrical Engineering from the University of Southern California. Her area of research is centered around the concept of humanized intelligence, the process of embedding human cognitive capability into the control path of autonomous systems. From 1993-2005, Dr. Howard was at NASA's Jet Propulsion Laboratory, California Institute of Technology. Following this, she joined Georgia Tech in July 2005 and founded the Human-Automation Systems Lab. She also serves as the Associate Director of Research for the GT Institute for Robotics and Intelligent Machines.