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

Paul Robinette, *Member, IEEE*, Ayanna M. Howard, *Senior Member, IEEE*, and Alan R. Wagner, *Member, IEEE*

***Abstract*—Robots have the potential to save lives in high-risk situations, such as emergency evacuations. To realize this potential, we must understand how factors such as the robot's performance, the riskiness of the situation, and the evacuee's motivation influence his or her decision to follow a robot. In this paper, we developed a set of experiments that tasked individuals with navigating a virtual maze using different methods to simulate an evacuation. Participants chose whether or not to use the robot for guidance in each of two separate navigation rounds. The robot performed poorly in two of the three conditions. The participant's decision to use the robot and self-reported trust in the robot served as dependent measures. A 53% drop in self-reported trust was found when the robot performs poorly. Self-reports of trust were strongly correlated with the decision to use the robot for guidance ($\phi(90) = +0.745$). We conclude that a mistake made by a robot will cause a person to have a significantly lower level of trust in it in later interactions.**

***Index Terms*—Cooperative systems, emergency guidance robot, human–robot interaction, human–robot trust.**

I. Introduction

TRUST is a requirement in every interaction that involves risk, from daily tasks to life-and-death situations. Emergency evacuations are dangerous situations with the potential for considerable loss of life, such as in the Station Nightclub Fire of 2003 [1]. Robots could potentially aid evacuees during these high-risk situations by providing safe guidance to exits. Ideally, robots stationed in buildings could provide immediate information in the form of accurate directions to unblocked exits.

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 robots to autonomously aid humans in high-risk tasks, humans must first trust the robots to perform high-risk tasks correctly. Exploring the conditions that result in a decision by a person to trust or not trust a robot in a high-risk, time-critical situation is a critical step toward developing acceptable emergency guidance robots.

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 during a high-risk, time-critical situation 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 to find the exit of a maze 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. We vary the behavior of the robot in the first round to determine the effect of successful and unsuccessful guidance on the participant's second choice. Two different methods were used to add time pressure to the scenario: a monetary bonus for a quick exit and a survival risk for not evacuating within a specified time.

The remainder of this paper begins by discussing work related to these concepts then outlines our conceptualization of trust followed by a description of our methodology. After this, we describe the setup and results from two separate experiments. Finally, we provide overall conclusions based on these experiments and some thoughts on future work.

II. Related Work

The general topic of human–robot trust is an active area of research. Several studies have examined the factors that influence trust in automation [2], [3], artificial agents [4], and robots [5]. Hancock *et al*. performed a meta-analysis over the existing human–robot trust literature identifying 11 relevant research articles and found that, for these papers, robot performance is most strongly associated with trust ($r = +0.34$). The research presented here confirms that robot performance is an important factor related to trust, but also explores the impact of environmental and motivational factors during an emergency evacuation scenario. Emergency evacuation scenarios are unique in that we

Manuscript received March 4, 2015; revised August 31, 2015, April 14, 2016, and August 4, 2016; accepted November 19, 2016. Date of publication January 20, 2017; date of current version July 13, 2017. This work was supported in part by the Air Force Office of Sponsored Research under Grant FA95501310169 and in part by the Motorola Foundation Professorship. This paper was recommended by Associate Editor L. D. Riek.

P. Robinette and A. R. Wagner are with the Georgia Tech Research Institute, Atlanta, GA 30332 USA (e-mail: probinette3@gatech.edu; alan.wagner@gtri.gatech.edu).

A. M. Howard is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ayanna.howard@ece.gatech.edu).

This paper has supplementary downloadable materials available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2017.2648849

cannot assume that the person will have experience with the robot.

Other related research has focused on the factors that affect trust in a robot [6]. Carlson *et al*. finds that reliability and reputation impact trust in surveys of how people view robots. In contrast to surveys, we use immersive simulations to record the person's actual behavior during an interaction involving trust. We also focus on initial interactions with a robot, rather than trust that has been built over a long history.

Desai *et al.* performed several experiments related to human–robot trust [7], [8]. This group's work primarily focused on the impact of robot reliability on a user's decision to interrupt autonomous operation by the robot. This is a qualitatively different question than the ones examined in this paper. In contrast to the work by Desai *et al.*, our work and the emergency evacuation scenario we investigate does not afford an opportunity for the human to take control of the robot. Instead, we are examining situations when people must choose to either follow the guidance of a robot or not. While this still explores the level of trust a person is willing to place in an autonomous robot, we believe that the difference between an operator's perspective on trust and an evacuee's perspective on trust is significant. The evacuee cannot affect the robot in any way and most choose between his or her own intuition and the robot's instructions.

Mason *et al*. use a maze environment to explore the impact of robot reliability on participant decisions to follow the robot [9]. Many of the study's findings, however, are inconclusive. Although their work bears some conceptual similarities, the research we present here is more focused on investigating the impact of trust on a person's decision-making during high-risk situations, such as emergency evacuation.

Researchers have also examined a human's decision to follow a robot's directions. Bainbridge *et al*. found that participants were likely to follow odd and potentially destructive instructions from a robot under certain conditions [10]. Our research does not examine odd or destructive instructions, but does investigate the factors that influence a person's decision to follow instructions from a robot in an emergency situation.

Experiments have shown that the nature of a robot's request can significantly influence the person's willingness to comply. Salem *et al.* performed an experiment to determine the effect of robot errors on unusual requests [11]. They found that participants still completed the odd request made by the robot in spite of errors. Our previous research also indicates that an individual will tend to follow an emergency evacuation robot's directions even when the person has no prior experience with a robot [12], [13]. In related prior research, we have also explored various situations where emergency guidance robots could improve human survivability during an evacuation [14]–[16]. This work includes developing robotic platforms that are capable of communicating intelligible guidance instructions to humans [17].

III. Conceptualizing Trust

Numerous researchers have proposed conceptions of trust that range from computational implementations of cognitive processes [18], to neurological changes in reciprocity games [19], to a probability of an agent performing a particular action [20]. Other researchers consider trust to have multiple forms, depending on the actors and environment [21]. 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* [2]. 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" [22]. The vast majority of trust research supports the connection between trust and risk [13], [20], [23]–[26] although not all researchers agree [27]. We are interested in human reactions to robot guidance in high-risk scenarios. We call this trust based on the above existing definitions that relate trust to risk.

There are a number of methodological challenges to exploring trust. A researcher must put a person at risk while also minimizing the potential for harm. Moreover, the participant must also believe that they are at risk and value the item at risk. For example, asking a person to invest $10 is a riskier decision for a destitute college student than for a wealthy individual. In developing the navigation task, we sought to generate a situation in which a person is placed 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. Because the majority of the trust literature uses the possible loss of money as the source of risk [19], [28], [29], we compared monetary motivation to an emergency motivation scenario.

To gauge the impact of the scenario on trust we need to be able to measure trust. Measurements of trust tend to focus on either self-reports, behavioral measures, or both. Desai *et al*. asked participants to self-report changes in trust [8]. Salem *et al*. equated trust to compliance with a robot's suggestions [11]. Measurements of the frequency of operator intervention in an otherwise autonomous system have also been used [30]. Our study examines both self-reports and behavioral measures in order to better understand if and when these measures correlate. We hypothesize that: *(H1) in a scenario that requires trust, participants will both self-report trust and make decisions that rely on the robot.* If participants choose to rely on the robot but do not report trust, then they may be motivated by some other aspect of the experiment (the novelty of following a robot for example). On the other hand, if participants report trust but tend not to rely on the robot then this result may be an indication of social desirability bias [31].

To investigate how a robot's initial performance impacts a person's trust during an emergency situation, we measure the change in trust while varying the robot's guidance performance. We hypothesize that: *(H2) participant's self-reported trust significantly decreases after the robot performs poorly.* Understanding the impact of a single guidance failure on the person and quantifying the disuse that results from that failure will be important for shaping future human–robot collaborative systems.

Finally, there are many ways for a robot to fail during an emergency. In the worst case, the use of emergency robot guides could result in increased fatalities. Our intent is to better understand how people react to a failed or failing robot and to understand how this experience influences their subsequent trust in a robot. We, therefore, created a number of plausible guidance robot failure modes. One type of failure is for the robot to not lead the person to the exit. We implemented this type of failure by having the robot stop in the maze at a location with no exit in sight. Another failure mode is for the robot to provide inefficient or circuitous guidance. This occurs when the robot successfully leads the person to the exit, but requires a great deal of time to do so. We hypothesize that: *(H3) self-reported trust is lower for participants guided by a robot that provides incorrect guidance than participants that are guided by a circuitous, inefficient robot.*

IV. Methods

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 were collected and used to confirm or refute the hypotheses presented above.

A. Participant Inclusion and Exclusion Criteria

Emergency guidance robots could potentially aid a large variety of people. In order to gather such a large variety of participants, crowdsourcing (via Amazon's Mechanical Turk service) was used to collect data for both experiments. Crowdsourcing is a method for collecting data from a relatively large, diverse set of people [32]. Crowdsourcing sites, such as 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. Additionally, 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. For experiment 1, 106 participants were recruited. For experiment 2, 129 participants were recruited. In each experiment, participants were recruited until 30 participants experienced each robot under each condition in the initial round.

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 [17]). A participant's data was excluded if comments were missing, nonsensical (e.g., if the comments were not understandable), or repeated throughout. Human participation in our experiments was approved by the Georgia Tech Institutional Review Board.

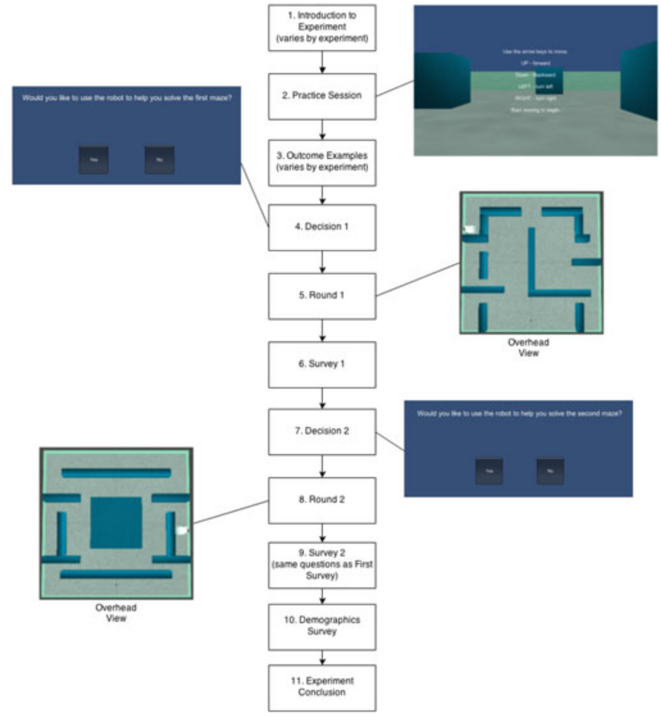

Fig. 1. Experimental protocol with screenshots from experiment. The entire experiment was presented in a Unity 3D Web Game, including the survey questions. Survey questions are described in Table I and Section IV-B.

B. Experimental Protocol

The same general experimental setup was used for both experiments (see Fig. 1). Participants began each experiment by accepting a 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 human–robot performances in the maze. Due to differing incentives, the nature of these examples varied with respect to the experiment. Examples are described further in Sections V and VI. The intent of the examples was to show participants that this is a scenario that requires trust, similar to our previous work [13]. 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 (see Table I). The survey collected 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. They were then offered another opportunity to decide if they wanted to use the guidance robot in the second round. Next, they navigated the maze in the second round and completed a short survey about their second round decision. Unknown to

TABLE I
SURVEY PRESENTED TO PARTICIPANTS AFTER EACH ROUND

Question
1. Did you choose to use the robot in the previous experiment?
2. Did you trust the robot?
3. Did you believe that the robot would find the exit quickly?
4. Were you motivated to find the exit as quickly as possible?
5. My decision to use the robot shows that I trusted the robot.

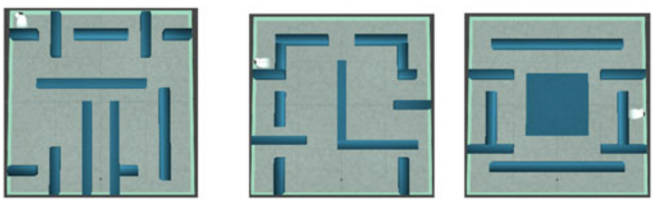

Fig. 2. Overhead views of the three environments used in both experiments. Environments were designed to be similar to office layouts. Corridors and rooms were used to give maze-like qualities to make the simulation challenging.

TABLE II
FAILED ROBOT GUIDANCE BEHAVIORS THAT WERE USED DURING A PILOT STUDY

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 min.
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.

the participant, 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 about participant age, gender, country of residence, occupation, and education level. This survey also asked if participants have worked with a real robot before.

Because the purpose of this research is to better understand how people react in an emergency situation, a simulation environment was created to resemble an office building. This environment included corridors and rooms designed to give it a maze-like appearance (see Fig. 2). Participants were placed in the environment with no prior experience and required to find a single exit.

C. Measuring Trust

The decision to use the robot was viewed as an indicator of trust. This decision served as a binary behavioral measure of trust: either the person trusts and uses the robot or the person does not trust and use the robot. Our conceptualization of trust focuses on the risk a person accepts when choosing to depend on the robot. Hence, we believed that the person's decision to use or not use the robot could serve as a measure of trust. In the first round, the participant must choose based on very little information, but in the second round the participant bases their decision on the robot's previous behavior. Thus, we felt that measuring the participant's decision to use or not use guidance from the robot at the beginning of the second round would provide a measure of their trust in the robot.

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 decision to follow or not follow the robot." In addition to the options to agree or disagree, we also offered the option of choosing "Trust was not involved in my decision." In pilot studies, we found that some participants felt that disagreeing with the trust statement meant that they actively distrusted the robot. We, therefore, provided a third option that clearly indicates they neither trust nor distrust the robot. Our results are based on people who answered that they trusted the robot. The use of binary measures as a supplement to or a replacement of Likert scores is common in trust research (e.g., [9], [33], [34]) and, we feel, more accurately reflects the types of high-risk decisions a person must make during an emergency.

D. Robot Behavior

The actions of the robot inform the human of the robot's ability to be trusted in future interactions. H2 examines how the robot's behavior affects the participants' self-reports of trust in the second round. H3 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 these failure modes were eliminated because participants were unable to determine that the robot had failed and hence resulted in an extremely long experiment completion time (see Table II for a listing of the robot guidance failure types that were not included in these experiments). Overall, three robot behaviors were defined that were used in the experiments.

1) Efficient navigation: the robot proceeds directly to the exit location (see Fig. 3). Robots that acted in this manner are capable of finding the exit within 30 s.
2) Circuitous navigation: the robot explores many possible routes before eventually finding the exit (see Fig. 3). Robots that acted in this manner are capable of finding the exit in 90 s.
3) Incorrect 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 30 s at a point at least 30 s from the exit.

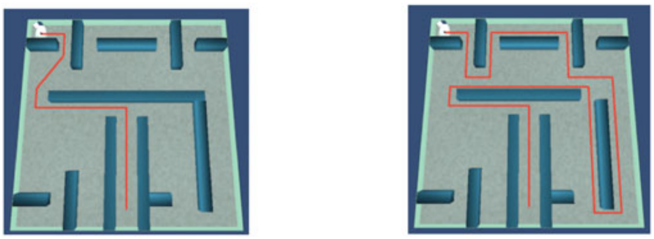

Fig. 3. Examples of efficient robot guidance (left) and circuitous robot guidance (right). During efficient guidance the robot knows exactly where the exit is and effectively mitigates the participant's risk. During circuitous guidance the robot searches for the exit, eventually finding it.

The robot followed a predefined set of waypoints throughout the environment to perform the behaviors. Waypoints were set near corners or occlusion points so that each was in view of the waypoint before it. The robot waited at each waypoint for the participant to approach before it moved to the next waypoint. The robot was allowed to move considerably faster than the participant so that it would always be leading. The exact time to reach each end point depended on the particular environment and on the participant.

V. Experiment 1: Bonus Scenario

The first experiment examines the use of losing money as a way to put participants at risk. We used the risk of losing a potential allotted bonus as the source of risk motivating participants' trust decisions. This is an established procedure in the trust literature [19], [28], [29]. Subjects were offered a $1 bonus if they could find the exit of a maze within 30 s. After the first 30 s 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.

The type of robot behavior (efficient, circuitous, incorrect) witnessed by the person served as the independent variable for this study. Measurements of trust served as the dependent variable. Both behavioral measures of trust and postround self-reports were collected. The correlation between these two measures was used to evaluate $H1$. Hypotheses $H2$ and $H3$ were examined by comparing trust measures between subjects that interacted with different types of robot guidance behaviors.

A. Experimental Setup

Although the experiments followed the same general procedure described in Fig. 1, some screens and text were unique for each experiment. This experimental setup can be seen in the video "Monetary Maze" in the supplemental material.

The first screen seen by the participants gave instructions. The simulated environments were specifically referred to as "mazes" to give the participant an idea of their complexity and goal. For this experiment, the robot displayed during the introduction and used in the rounds was a Willow Garage TurtleBot 2. The three-dimensional model of the robot was created out of CAD files distributed by the manufacturer.

After the practice session, the participants were informed of the performance-based bonus and how to obtain it. Participants were given three example performances (see Fig. 4 bottom left) for the navigation task:

1) stated "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;
2) stated "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; and
3) stated "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.

For this experiment, at the beginning of each round participants were informed that their bonus was currently set at $1.00 (see Fig. 4 top right). When the participant began moving, a timer in the top left of the screen displayed the time spent navigating to a tenth of a second precision. The bonus was prominently displayed in the top right corner. After 30 s of navigating the maze, the bonus began to decrease at a rate of $0.0167 per second (see Fig. 4 bottom right). The bonus was completely depleted after 90 s. 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 Section IV.

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 the experiment.

In this experiment, we also asked one additional survey question in order to better understand choices for following the robot. Participants were asked to rate their motivations with respect to time, money, and enjoyment on a seven point Likert scale. They were then asked to rank these motivations in terms of importance from most to least important. The additional survey question was only included to help design better experiments.

B. Results

A total of 106 participants (mean age $= 31.0$, standard deviation $= 8.4$, 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. Fig. 5 depicts the number of participants who used the robot in rounds 1 and 2 for the efficient and circuitous/incorrect 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%, $\chi^2(1,\ N = 60) = 68.76,\ p < 0.001$) when the participants experience a circuitous or incorrect robot in the first round. Only a 4% ($\chi^2(1,\ N = 30) = 0.11,\ p = 0.739$) decrease in trust was reported by participants that were guided by an efficient robot. There is a 40% difference in the level of trust in the second round between the efficient guidance behavior and the circuitous/incorrect behaviors ($\chi^2(1,\ N = 90) = 12.85$, $p < 0.001$). Efficient robots saw a 17% drop in decision to use

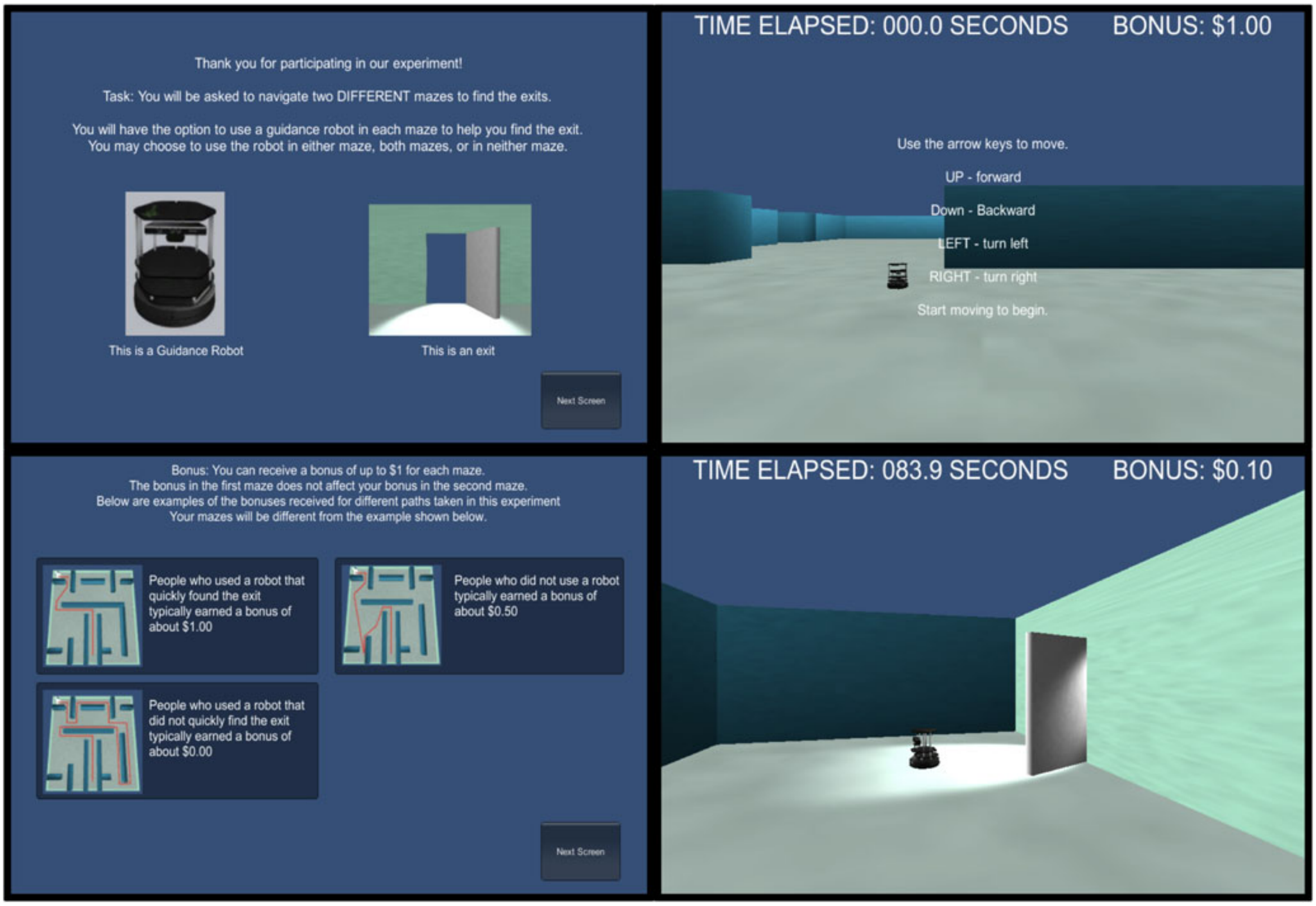

Fig. 4. Screenshots from the bonus scenario experiment. The figure depicts the introduction screen (top left), example outcomes (bottom left), beginning of a round (top right), and successful navigation to an exit (bottom right).

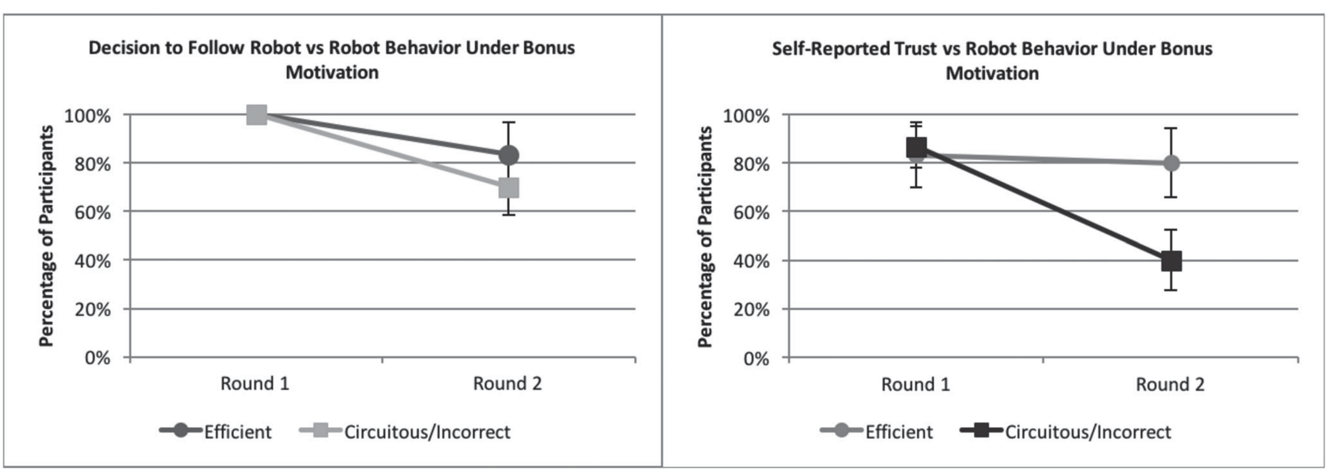

Fig. 5. Change in decision to use robot (left) and self-reported trust (right) between the two rounds for the successful and unsuccessful robots. Note that a majority of participants continued to use the circuitous/incorrect robots even though half had lost their trust in the robot. Error bars represent 95% confidence intervals.

the robot between the two rounds ($\chi^2(1,\ N = 30) = 5.45,\ p = 0.020$) while circuitous and incorrect robots dropped 30% between the two rounds ($\chi^2(1,\ N = 60) = 42.33,\ p < 0.001$). Still, there was only a 13% difference between the efficient robot usage and the other robots usage in the second round ($\chi^2(1,\ N = 90) = 1.87,\ p = 0.172$). Fig. 6 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 follow-

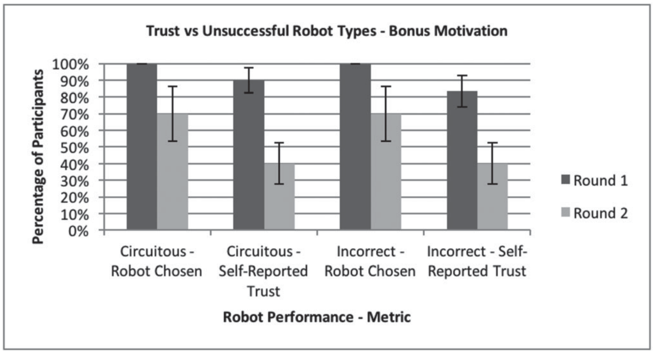

Fig. 6. Change in decision to use robot and self-reported trust between the two rounds for the circuitous and incorrect robots. The same number of participants chose to use each and the same number reported trust in each in the second round. Error bars represent 95% confidence intervals.

TABLE III
SUMMARY OF COMMENTS FROM EXPERIMENT 1

Robot Behavior	Used Robot?	Self-Reported Trust	Comment Description
Efficient ($n = 30$)	Yes ($n = 25$)	Positive ($n = 22$)	Robot performed well ($n = 21$)
			Did not trust robot, trusted programmers ($n = 1$)
		Neg./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 ($n = 2$)
		Neg./Neutral ($n = 3$)	No complaint about robot, wanted to try experiment for themselves ($n = 1$)
			Thought robot would perform worse in second round ($n = 1$)
Circuit. ($n = 30$)	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 ($n = 1$)
		Neg./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 ($n = 1$)
		Neg./Neutral ($n = 8$)	Robot performed poorly ($n = 7$)
			No complaint about robot, wanted to try experiment for themselves ($n = 1$)
Incorrect ($n = 30$)	Yes ($n = 21$)	Positive ($n = 11$)	Thought robot would perform better in second round ($n = 5$)
			Did not realize robot performed poorly ($n = 3$)
			Curiosity ($n = 3$)
		Neg./Neutral ($n = 10$)	Curiosity ($n = 6$)
			Robot performed better than human alone ($n = 1$)
	No ($n = 9$)	Positive ($n = 1$)	Unclear response ($n = 1$)
		Neg./Neutral ($n = 8$)	Robot performed poorly ($n = 8$)

ing the robot and reporting trust in the robot, $\phi(106) = +0.628$ for round 1 and $\phi(90) = +0.422$ for round 2.

We examined the survey comments to better understand each participant's rationale. Table III summarizes the most common comments from round 2. Note that, of the people that were guided by a circuitous or incorrect 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.

We performed an analysis on our motivational survey to better understand the participants. About half of the 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 chose 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.

C. Discussion

The results from this experiment highlight the methodological challenges associated with investigating human–robot trust. The data indicate a large, significant decrease in self-reported trust when the robot fails compared to when the robot does not fail (53%). Yet, the results also indicate only a small, insignificant difference (13%) in the person's decision to use the robot after a failure compared to after a successful interaction. Participants in this scenario are essentially stating that they will use the robot even though they do not trust it. This contradicts our own intuition as well as related work, such as [7] and [8], who found that operators typically stopped using autonomous modes on robots that performed poorly.

The comments highlight some of the methodological challenges of human–robot trust research. Risk is a major component of our definition of trust [13], [20], [22]–[25]. Characteristics of the experimental scenario can influence a subject's perceived risk differently. From an empirical point of view, we would like to control the factors that influence the subject's perceived risk. We used monetary incentives in this experiment to maintain continuity with common methods for putting a person at risk in order to explore trust [19], [28], [29]. Our results appear to indicate that the use of monetary incentives in high-risk trust research may be methodologically flawed.

These results 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." Thus, instead of a monetary risk, participants experienced a survival risk.

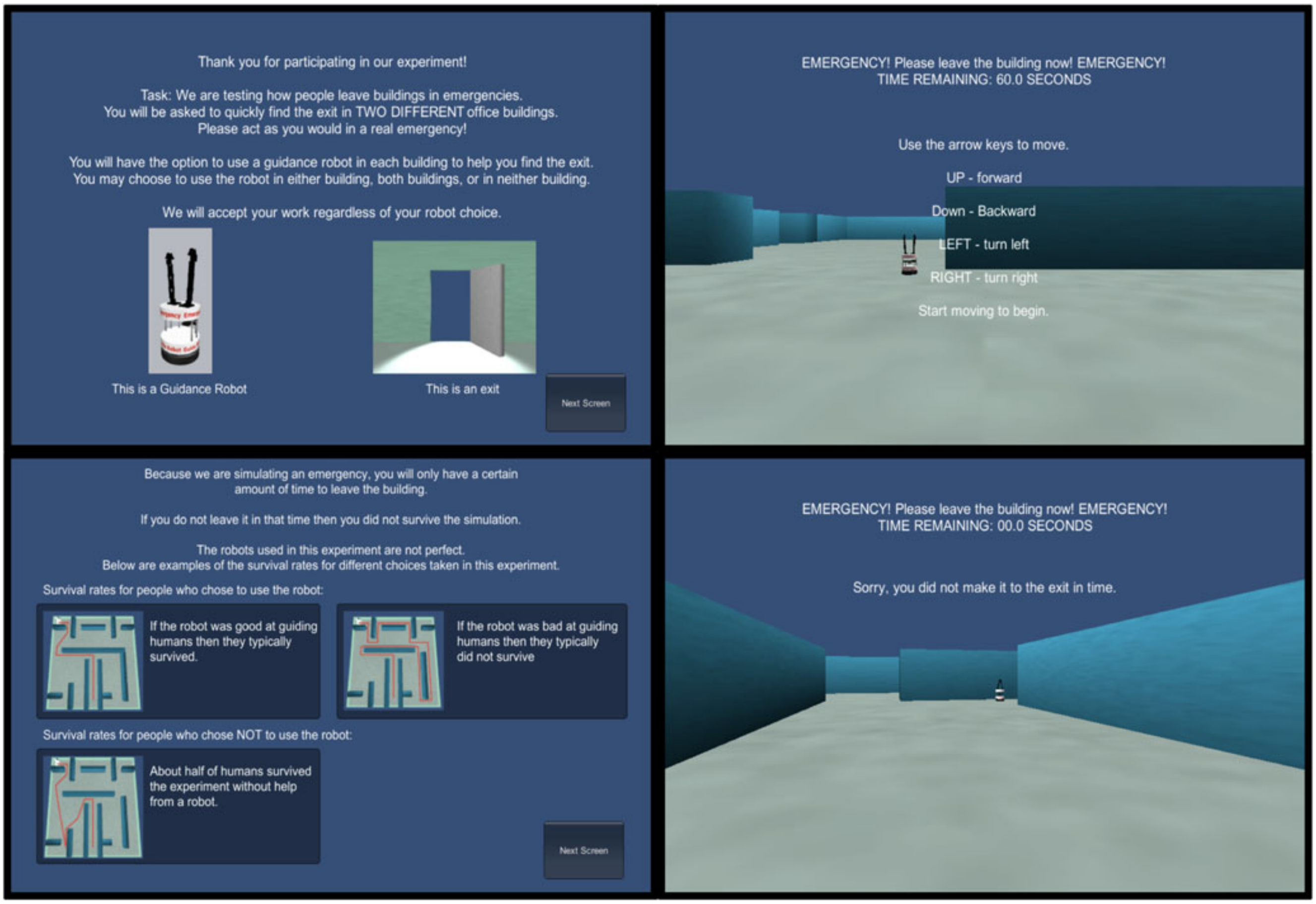

Fig. 7. Introduction screen for the emergency scenario experiment is depicted in the top left. Note that the robot is different from in experiment 1. Additionally, participants were told that this experiment was to determine how humans evacuate buildings. The screen on the bottom left depicts example results. Participants were shown overhead views of the example environment with survival possibilities. The screen on the top right presents the beginning of the first round of the experiment. The timer counted down and was moved to the center of the screen for maximum visibility. Text indicated that an emergency had occurred. An example of an unsuccessful exit is presented in the bottom right. Text informed the participant there was no time remaining. The robot can be seen in the distance.

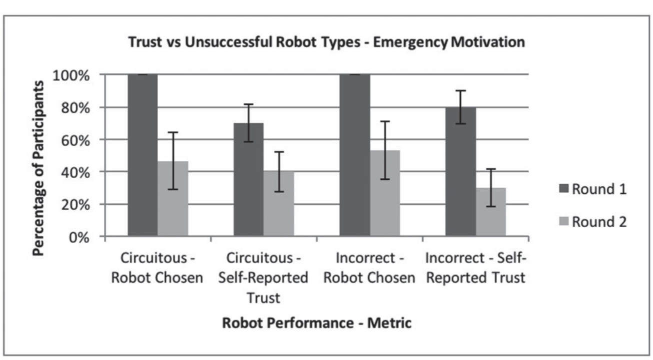

Fig. 8. Change in decision to use robot (left) and self-reported trust (right) between the two rounds for efficient and circuitous/incorrect robots. Note that the decision to use the robot dropped with self-reported trust in this experiment, unlike in experiment 1. Error bars represent 95% confidence intervals.

VI. Experiment 2: Emergency Scenario

In a second 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 both 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.

A. Experimental Setup

There were several differences between this experiment and experiment 1. First, the introduction screen stated "We are testing how people leave a building in emergencies" and asked them to "Please act as you would in a real emergency!" (see Fig. 7 top left). The word "building" was used instead of "maze" to further reinforce the emergency portion of the simulation. This experimental setup can be seen in the video "Emergency Maze" in the supplementary material available at http://ieeexplore.ieee.org.

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'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 [17].

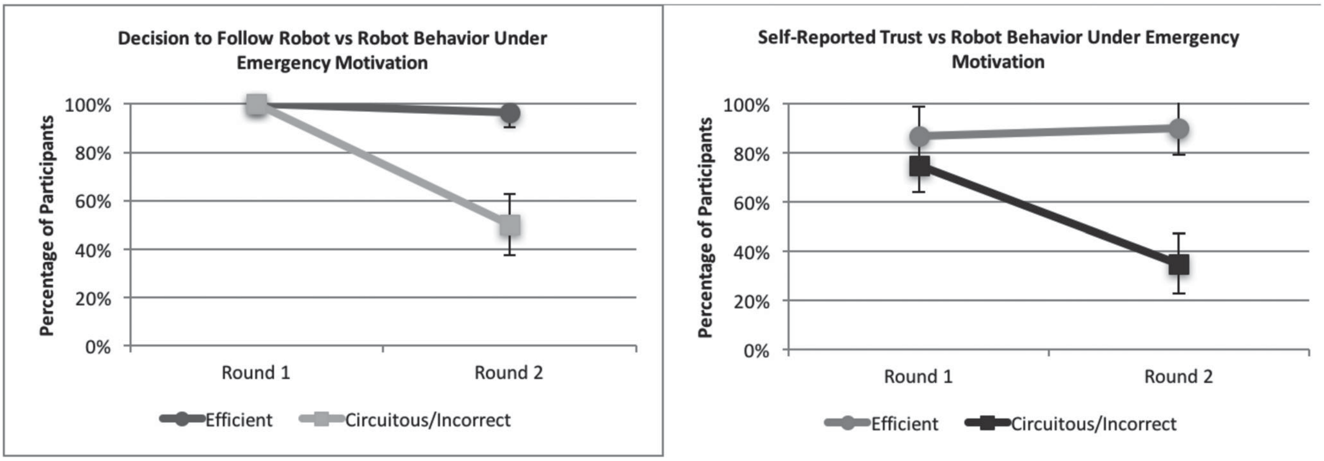

Fig. 9. Change in decision to use robot and self-reported trust between the two rounds for the circuitous and incorrect robots. While the results are not identical in this round, as they were in experiment 1, they are still not statistically significant. Error bars represent 95% confidence intervals.

For this experiment, each round ended after 60 s 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 performances (see Fig. 7 bottom right):

1) stated "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;
2) stated "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; and
3) stated "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.

During each round, the words "EMERGENCY! Please leave the building now! EMERGENCY!" appeared as well as the time remaining to exit (to a tenth of a second precision) in the top-center of the participants' view throughout the entire round (see Fig. 7 top right).

Other than these changes, both experiments were identical. Participants were again required to complete the same survey examining their trust in the robot and reasoning for choosing the robot for both rounds.

B. Results

A total of 129 participants (mean age = 31.8, standard deviation = 8.4, 60.5% male) completed the second experiment, 69.8% of which decided to use the robot in the first round. As shown in Fig. 8, the decision to follow the robot decreases by 50% in the second round when the participant interacts with a circuitous/incorrect robot in the first round ($\chi^2(1,\ N = 60) < 0.01,\ p < 0.001$), compared to just 3% when an efficient robot is used first ($\chi^2(1,\ N = 30) = 1.02,\ p =\ 0.313$). There was a 47% difference in usage between the efficient guidance behavior and the failed behaviors ($\chi^2(1,\ N = 90) = 19.29,\ p < 0.001$). Self-reported trust follows a similar trend with trust decreasing 53% when participants experienced a circuitous/incorrect robot ($\chi^2(1,\ N = 60) < 0.01,\ p < 0.001$) and self-reported trust increasing by 3% after interacting with an efficient robot in the first round ($\chi^2(1,\ N = 30) = 0.16,\ p =\ 0.688$). There was a 55% difference between efficient robot and failed robot trust levels in the second round ($\chi^2(1,\ N = 90) = 24.31,\ p < 0.001$). Fig. 9 shows the results for the different failure modes. The type of failure had minimal impact in the participant's decision to follow ($\chi^2(1,\ N = 60) = 0.27,\ p = 0.606$). There was also a minimal change in self-reported trust $\chi^2(1,\ N = 60) = 1.15,\ p = 0.284$). 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.

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 IV. Note that not all participants' comments are included in this table for brevity and some participants gave multiple reasons for their actions.

C. Discussion

The results show that in this case, in contrast to the bonus scenario, both self-reported trust and the decision to follow the robot significantly decrease after a failure. 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. Moreover, a failure caused a 53% decrease in self-reported trust. These results support our second hypothesis (H2). Additionally, the results agree with existing literature about operator–robot trust [7], [8]. Notably, these results do not agree with studies in which robots asked participants to perform unusual requests [10], [11]. We believe this is because our participants were in a scenario that simulated survival risk and thus were focused on evacuating quickly; however, it should be noted that our experiments were conducted in a virtual environment while Bainbridge *et al.* and Salem *et al.* performed their experiments in a real-world environment.

Self-reports of trust and the decision to use the robot were strongly correlated. Also, possibly indicative of perceived risk,

TABLE IV
SUMMARY OF COMMENTS FROM EXPERIMENT 2

Robot Beh.	Follow Dec.	Trust Answer	Comment Description
Efficient ($n = 30$)	Yes ($n = 29$)	Positive ($n = 27$)	Robot performed well ($n = 24$)
		Neg./ 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 ($n = 0$)	
		Neg./ Neutral ($n = 1$)	Thought robot would perform worse in second round ($n = 1$)
Circuit. ($n = 30$)	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 ($n = 2$)
		Neg./ Neutral ($n = 3$)	Curiosity ($n = 3$)
	No ($n = 15$)	Positive ($n = 1$)	Trusted robot to NOT find exit ($n = 1$)
		Negative/ Neutral ($n = 14$)	Robot performed poorly ($n = 13$)
			No complaint about robot, wanted to try experiment for themselves ($n = 2$)
Incorrect ($n = 30$)	Yes ($n = 16$)	Positive ($n = 9$)	Robot performed better than human alone ($n = 6$)
			Thought robot would perform better in second round ($n = 3$)
		Neg./ Neutral ($n = 7$)	Curiosity ($n = 5$)
			Robot performed better than human alone ($n = 2$)
	No ($n = 14$)	Positive ($n = 0$)	
		Neg./ Neutral ($n = 14$)	Robot performed poorly ($n = 12$)
			No complaint about robot, wanted to try experiment for themselves ($n = 2$)

15.1% fewer people followed the robot in round 1. While a majority still chose to use the robot, 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 perceived survival risk). The results from this scenario support hypothesis H1.

The comments also 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. This data serve 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 circuitous robot versus an incorrect robot that stopped moving before arriving at the exit. It appears they do not discriminate based on how the robot failed, only that it did fail. The results do not, therefore, support hypothesis H3.

VII. CONCLUSION AND FUTURE WORK

This paper has explored how a person's trust in a guidance robot during an emergency evacuation is affected by poor or failed navigation advice.

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 low risk situations (monetary bonus) people may act as if they trust the robot after a failure even if they self-report little trust. We found that in higher risk scenarios (simulated emergency) participant's self-reports matched their decision to use the robot. Experiments which attempt to equate the person's risk to a bonus appear to underestimate other motivations such as time and fun.

There are a number of practical and theoretical implications of this work. From a methodological stand point, this work could influence the way trust in robots is investigated. Moreover, our work shows that a majority of people are willing to accept guidance information from a completely unknown robot but that a failure by the robot sharply reduces trust in the system. Whether or not this reduction in trust results in disuse appears to relate to the risk entailed by the situation. With respect to the development of emergency evacuation robots, our results should influence how these robots are designed and how they communicate failure states. Overall, this work begins to explain how and why a robot's behavior impacts a person's trust. Hence, we believe that these findings are broadly applicable to a large variety of human–machine systems.

This work is not without limitations. The results are likely influenced by the fact that the study utilized 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 the first experiment and to "survive" in the second experiment, we cannot be sure that their decision-making in these simulated environments matches what their behavior would be in a real emergency. Still, the use of an internet-based simulation also means that these results are not based on a small sample and that the subject population is broad and diverse compared to traditional laboratory experiments dominated by college students.

Our more recent work has examined similar situations in real-world simulations of emergencies and has shown that people tend to overtrust robots, even if the robots have previously committed an error [35]. This raises questions about why participants behave differently in the two different paradigms. It is challenging to realistically simulate an emergency, either in a virtual or a physical environment, so it is difficult to determine which environment produces a response from participants that is most similar to people's real-world reactions to robots during actual emergencies.

Hence, 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. In more recent work, we performed an experiment in which a robot repaired trust in a virtual emergency scenario by apologizing at the proper time [36]. Performing a similar experiment in a real-world experiment, similar to [9], could help us understand how a robot's apology influences a person's behavior. This paper specifically explores trust decisions at discrete points; however, trust in these situations tends to be a continuous decision. Emergency evacuation experiments which track a person's trust in the system continuously as it fails and attempts to recover trust, similar to [8], could provide additional insight.

As robots enter everyday life, we must be conscious of their effect 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.

REFERENCES

[1] W. Grosshandler, N. Bryner, D. Madrzykowski, and K. Kuntz, "Report of the technical investigation of the station nightclub fire," *NIST NCSTAR*, vol. 1, no. 2, Jun. 2005.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, pp. 50–80, 2004.

[3] K. A. Hoff and M. Bashir, "Trust in automation integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[4] J. Sabater and C. Sierra, "Review of computational trust and reputation models," *Artif. Intell. Rev.*, vol. 24, pp. 33–60, 2005.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[6] M. S. Carlson, M. Desai, J. L. K. H. Drury, and H. A. Yanco, "Identifying factors that influence trust in automated cars and medical diagnosis systems," in *Proc. AAAI Spring Symp. Intersection Robust Intell. Trust Auton. Syst.*, 2014, pp. 20–27.

[7] M. Desai *et al.*, "Effects of changing reliability on trust of robot systems," in *Proc. 2012 7th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2012, pp. 73–80.

[8] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2013, pp. 251–258.

[9] E. Mason, A. Nagabandi, A. Steinfeld, and C. Bruggeman, "Trust during robot-assisted navigation," in *Proc. 2013 AAAI Spring Symp. Series*, 2013, pp. 54–59.

[10] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *Int. J. Soc. Robot.*, vol. 3, no. 1, pp. 41–52, 2011.

[11] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human–robot cooperation and trust," in *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2015, pp. 141–148.

[12] P. Robinette and A. M. Howard, "Trust in emergency evacuation robots," in *Proc. 2012 IEEE Int. Symp. Saf., Secur., Rescue Robot.*, 2012, pp. 1–6.

[13] A. R. Wagner and P. Robinette, "Towards robots that trust: Human subject validation of the situational conditions for trust," *Interact. Stud.*, vol. 1, no. 1, pp. 89–117, 2015.

[14] P. Robinette and A. M. Howard, "Incorporating a model of human panic behavior for robotic-based emergency evacuation," in *Proc. IEEE RO-MAN*, 2011, pp. 47–52.

[15] P. Robinette, P. A. Vela, and A. M. Howard, "Information propagation applied to robot-assisted evacuation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 856–861.

[16] P. Robinette, A. R. Wagner, and A. Howard, "Building and maintaining trust between humans and guidance robots in an emergency," in *Proc. AAAI Spring Symp.*, 2013, pp. 78–83.

[17] P. Robinette, A. R. Wagner, and A. M. Howard, "Assessment of robot guidance modalities conveying instructions to humans in emergency situations," in *Proc. 23rd IEEE Int. Symp. Robot Hum. Interact. Commun.*, 2014, pp. 1043–1049.

[18] C. Castelfranchi and R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*. New York, NY, USA: Wiley, 2010.

[19] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz, and P. R. Montague, "Getting to know you: Reputation and trust in two-person economic exchange," *Science*, no. 308, pp. 78–83, 2005.

[20] D. Gambetta, "Can we trust trust?," in *Trust, Making and Breaking Cooperative Relationships*, D. Gambetta, Ed. Oxford, U.K.: Blackwell, 1990, pp. 213–237.

[21] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," *Intell. Syst.*, vol. 28, no. 1, pp. 84–88, 2013.

[22] A. R. Wagner, The role of trust and relationships in human–robot social interaction, Ph.D. dissertation, School Interact. Comput., Georgia Inst. Technol., Atlanta, GA, USA, 2009.

[23] M. Deutsch, "Cooperation and trust: Some theoretical notes," in *Proc. Nebraska Symp. Motivation*. Lincoln, NB, USA, 1962, pp. 275–315.

[24] N. Luhmann, *Trust and Power*. Chichester, U.K.: Wiley, 1979.

[25] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Acad. Manage. Rev.*, vol. 23, pp. 393–404, 1998.

[26] J. K. Rilling, A. G. Sanfey, J. A. Aronson, L. E. Nystrom, and J. D. Cohen, "The neural correlates of theory of mind within interpersonal interactions," *NeuroImage*, vol. 22, no. 4, pp. 1694–1703, 2004.

[27] C. C. Eckel and R. K. Wilson, "Is trust a risky decision?" *J. Econ. Behav. Org.*, vol. 55, no. 4, pp. 447–465, 2004.

[28] R. Axelrod, *The Evolution of Cooperation*. New York, NY, USA: Basic Books, 1984.

[29] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games Econ. Behav.*, vol. 10, pp. 122–142, 1995.

[30] F. Gao, A. S. Clare, J. C. Macbeth, and M. L. Cummings, "Modeling the impact of operator trust on performance in multiple robot control," in *Proc. AAAI Spring Symp., Trust Auton. Syst.*, 2013, pp. 16–22.

[31] D. P. Crowne and D. Marlowe, "A new scale of social desirability independent of psychopathology," *J. Consulting Psychol.*, vol. 24, no. 4, pp. 349–354, 1960.

[32] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment Decis. Making*, vol. 5, no. 5, pp. 411–419, 2010.

[33] A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen, "The neural basis of economic decision-making in the ultimatum game," *Science*, vol. 300, no. 5626, pp. 1755–1758, 2003.

[34] J. A. Aimone, D. Houser, and B. Weber, "Neural signatures of betrayal aversion: An fMRI study of trust," *Proc. Roy. Soc. Lond. B, Biol. Sci.*, vol. 281, no. 1782, 2014, Art.ID. 20132127.

[35] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2016, pp. 101–108.

[36] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *Proc. Int. Conf. Soc. Robots*, 2015, pp. 574–584.

[37] P. Robinette, A. R. Wagner, and A. Howard, "Modeling human-robot trust in emergencies," in *Proc. AAAI Spring Symp.*, 2014, pp. 62–63.

[38] P. Robinette, A. R. Wagner, and A. M. Howard, “Investigating human-robot trust in emergency scenarios: methodological lessons learned,” in *Robust Intelligence and Trust in Autonomous Systems*. New York, NY, USA: Springer, 2016.

[39] M. E. Schweitzer, J. C. Hershey, and E. T. Bradlow, “Promises and lies: Restoring violated trust,” *Org. Behav. Hum. Decis. Process.*, vol. 101, no. 1, pp. 1–19, 2006.

Paul Robinette (M’06) received B.S. degrees in computer engineering and physics and the M.S. degree in computer engineering from the Missouri University of Science and Technology, Rolla, MO, USA, in 2008 and 2010, respectively, and the Ph.D. degree in robotics from the Georgia Institute of Technology, Atlanta, GA, USA, in December 2015.

He is currently a Research Engineer with the Georgia Tech Research Institute, Atlanta. His work focuses on human trust in guidance robots during emergency evacuations.

Ayanna M. Howard (S’92–M’99–SM’04) received the B.S. degree in engineering from Brown University, Providence, RI, USA, in 1993, and the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1994 and 1999, respectively.

She is the Linda J. and Mark C. Smith Chair in Bioengineering in the School of Electrical and Computer Engineering, Georgia Institute of Technology, and Chief Technology Officer of Zyrobotics, Atlanta, GA, USA, a Georgia Tech VentureLab startup company.

Alan R. Wagner (M’02) received the B.S. degree in psychology from Northwestern University, Evanston, IL, USA, in 1997, the M.S. degree in computer science from Boston University, Boston, MA, USA, in 2001, and the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta, GA, USA, in 2009.

He currently holds a position as a Senior Research Scientist with the Georgia Tech Research Institute and is a member of the Institute for Robotics and Intelligent Machines, Atlanta.