Overtrust of Robots in Emergency Evacuation Scenarios

Paul Robinette*†, Wenchen Li†, Robert Allen†, Ayanna M. Howard*, Alan R. Wagner†
*School of Electrical and Computer Engineering
Georgia Institute of Technology
†Georgia Tech Research Institute
Atlanta, GA, USA
probinette3@gatech.edu, wenchenli@gatech.edu, rallen35@gatech.edu, ayanna.howard@ece.gatech.edu, alan.wagner@gtri.gatech.edu

***Abstract*—Robots have the potential to save lives in emergency scenarios, but could have an equally disastrous effect if participants overtrust them. To explore this concept, we performed an experiment where a participant interacts with a robot in a non-emergency task to experience its behavior and then chooses whether to follow the robot's instructions in an emergency or not. Artificial smoke and fire alarms were used to add a sense of urgency. To our surprise, all 26 participants followed the robot in the emergency, despite half observing the same robot perform poorly in a navigation guidance task just minutes before. We performed additional exploratory studies investigating different failure modes. Even when the robot pointed to a dark room with no discernible exit the majority of people did not choose to safely exit the way they entered.**

I. Introduction

Fire emergencies are dangerous scenarios that require people to evacuate a building quickly. Evacuees have little time to make decisions or select optimal paths, so they rely on existing technology, such as emergency exit signs and evacuation maps, as well as information gleaned from authority figures, to find the best way out. As robots become more pervasive in everyday life, we can expect them to one-day guide evacuees in emergencies. There is considerable risk of injury or even death to evacuees in this situation, so we must understand the factors that affect human-robot trust in these scenarios before such robots are deployed.

Emergency guide robots have the potential to save human lives during fires. For example, the Station Nightclub in 2003 claimed about 100 lives in a matter of minutes [6]. Moreover, the number of environments demanding quick evacuation is growing. Globally, the number of buildings over 200 meters tall has increased from 21 in 1980 to 935 in 2014 [14]. These buildings demand quick, coordinated evacuation of hundreds or thousands of people. Ideally, in situ guide robots would autonomously assess the situation and efficiently lead victims to safety while providing information to first responders (see [11] for an example).

While there are a number of issues related to the building and control of an emergency evacuation robot, this paper focuses on the human-robot interaction aspects of such a system. In particular, this paper examines questions related to trust of such systems by evacuees. Given that evacuees will place their lives at risk, dependent on the actions of these robots, will people trust and follow the directions of the robot? Under what conditions will they stop following it and why?

In the next section, we discuss related work in the domain of human-robot trust. We then describe our methodology, including our concept of trust and trust measurement. Next, we introduce our experiment and its results. Based on these initial experiments, we then present some exploratory studies. This paper concludes with a discussion of the experiments and directions for future work.

II. Related Work

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 [3]. Salem et al. equated trust to compliance with a robot's suggestions [15]. Measurements of the frequency of operator intervention in an otherwise autonomous system have also been used [5]. Our study examines both self-reports and behavioral measures by recording the participant's decision to follow instructions from a robot in an emergency and then asking them if this decision meant that they trusted the robot.

Much of the research on human-robot trust has focused on the factors that underpin trust in a robot. 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 [7]. Desai et al. performed several experiments related to human-robot trust [3]. This group's work primarily focused on the impact of robot reliability on a user's decision to interrupt an autonomously operating robot. They found that poor robot performance negatively affected the operator's trust of the robot; however, 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 must choose between his or her own intuition and the robot's instructions.

In contrast to the work above, some researchers have found that participants will disregard prior experience with the robot when the robot asks them to perform an odd and potentially destructive task. Salem, et al. performed an experiment to determine the effect of robot errors on unusual requests [15]. They found that participants still completed the odd request made by the robot in spite of errors. Bainbridge et al. found that participants were willing to throw books in the trash when a physically present robot gave the instruction, but not when the robot was located in another room communicating through a video interface [1]. This experiment did not expose participants to any robot behavior prior to the instructions. Our experiment is designed to put participants under time pressure to make a potentially risky decision. We hypothesize that this situation would produce a different effect than Salem et al. and Bainbridge et al. found.

Emergency guide robots have demonstrated their usefulness when deployed as stationary audio beacons [16] and in simulations of large evacuations [11]. Other work has found that participants will generally follow a guide robot in a virtual emergency simulated with a 3D game engine even if they have no prior experience with the robot [17]. This trust drops after a participant experiences a robot that performs its task poorly in the same simulator [13]. In a similar experiment conducted in a virtual environment [10], a robot guided people to a meeting room in a non-emergency interaction. In one condition, the robot led them directly to the meeting room and in the other, the robot took a circuitous route. Participants were then alerted to an emergency and given the choice to follow the robot or find their own way out. The researchers found that significantly fewer participants chose to follow the robot in the emergency if it initially took a circuitous route to the meeting room when compared to a robot that had taken an efficient route initially, even though the behavior in the emergency interaction was identical. Moreover, the researchers found that the timing of trust repair techniques had a significant effect on a person's decision to follow the robot in the emergency.

III. Methodology

In contrast to this prior work, we endeavored to investigate human-robot trust during high-risk situations. Unlike low risk situations, high-risk situations may engage fight-or-flight responses and other cognitive faculties which impact a person's trust in difficult to predict ways. To the best of our knowledge, this is the first attempt to examine human-robot trust in comparatively high-risk situations.

To create a high-risk situation, we utilize an emergency evacuation scenario in which a robot first guides a person to a meeting room. Next, we simulate an emergency using artificial smoke and smoke detectors and have the robot provide guidance to an exit. The participant is not informed that an emergency scenario will take place prior to hearing the smoke alarm. After exiting the meeting room, the participant has an opportunity to follow the guidance suggested by a robot to a previously unknown exit or, alternatively, to follow a lighted emergency exit sign and exit through the door they used to enter the building.

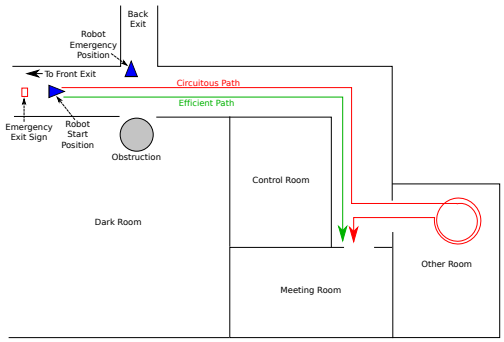

Figure 1. Layout of experiment area showing efficient and circuitous paths.

We record whether the participant uses guidance from the robot or returns to the main entrance to leave the building in the emergency. We supplement this measurement with survey questions. Prior studies of human behavior during evacuations have found that evacuees tend to exit through the main entrance of the building [2], [6], so participants would be biased towards exiting through the front exit if the robot was not present.

A. Robot Behaviors

Prior studies using 3D simulations of emergency guidance robots have found that people tend to trust robots that performed well in prior interactions but not trust robots that performed poorly in prior interactions [13], [10]. Inspired by this work, we use two behaviors (Figure 1) to bias participants for or against trusting the robot in a later emergency:

- Efficient: The robot takes the most direct path to its destination.
- Circuitous: While navigating to its destination, the robot enters an unrelated room and performs two circles before exiting and providing guidance to its destination.

B. Experimental Protocol

Participants began the experiment by reading and signing a consent form. Participants then completed Survey 1, which asked them to agree or disagree with ten statements about robots (e.g. "Robots are dangerous" and "Robots are helpful") and asked them to rate their current feelings on a seven point numbered Likert scale. Five feelings were included on the survey: comfortable, safe, relaxed, confused, and scared. Upon completing the survey, participants were given a device that would produce a loud sound in case they needed to abort the experiment for any reason, were introduced to the robot and instructed to follow it to a meeting room. They were not given any information about a potential emergency situation. They were told that the robot would inform them when they

had reached the meeting room by pointing with its arms (this gesture was previously evaluated in [12]). Participants were told that written instructions were waiting in the meeting room.

During this non-emergency guidance phase, participants followed the robot to the meeting room. The robot was remote controlled from a hidden location in order to ensure participant safety. The robot performed either efficient or circuitous guidance along the way. Participants were randomly assigned to each condition.

After arriving in the meeting room, participants followed written instructions labeled as "Meeting Room Instructions" and posted in two locations on the wall as well as on a table:

1) Sit down at the table.
2) Complete the survey on the clipboard.
3) Close the door to this room.
4) Read the article on the table. Mark important sections of the article and make any notes necessary on the page. You will be asked questions about this document after the experiment is completed.

This survey (Survey 2) first asked participants "Did the robot do a good job guiding you to the meeting room?" and why or why not. Then the same ten robot statements as in Survey 1 were asked. Finally, participants were asked how they felt (using the same five feelings as before) while moving through the building to the meeting room. The article on the meeting room table was an IEEE Spectrum article about indoor navigation technologies. It did not mention emergency guide robots and was chosen because it was somewhat related to the experiment so far, but would not bias participants towards or against the robot.

A timer on a smoke generator was triggered when the participant closed the meeting room door. This was supposed to occur after the participant finished the survey and before they started the article, but some participants closed the door early. Other participants never closed the door and an experimenter closed it when it was clear from camera footage that participants had completed the survey. The timer counted down for three minutes and then released artificial smoke into the hallway outside the meeting room door for twelve seconds. In tests, this was found to be sufficient to give a smoky appearance to the hallway outside the meeting room but not enough to impair participant vision such that they would injure themselves (see Figure 2 for example). The robot was placed at a hallway along the path to the main entrance and pointed down the hallway to what appeared to be another exit (Figure 1).

The emergency phase of the experiment began when the artificial smoke triggered a smoke detector. This could take approximately thirty to ninety seconds after the smoke stopped. Participants exited the room, proceeded down the hallway to the first corner, and then observed the robot. They then decided to either follow its guidance or proceed to the main entrance via the path they used to go to the meeting room.

An experimenter was waiting at the entrance and another was waiting at the back exit, where the robot was pointing, during the simulated emergency. When the participant had clearly made their choice by walking further down the hallway to the main entrance or down the hallway to the back exit, the closest experimenter stopped him or her and informed him or her that the emergency was a part of the experiment. The participant was then given a third survey, where they were asked about the realism of the emergency, the method they used to find an exit, whether their decision to use the robot indicated that they trusted it, the same ten statements as before, the five questions on feelings, and demographic information.

Figure 2. Example of smoke-filled hallway after smoke detector is triggered.

Aside from three experimenters and one participant, no one else was in the building at the time of the experiment. The study was performed in a double-blind manner in which neither the experimenters that interacted with the participants nor the participants themselves knew what type of behavior the robot employed. This experiment was approved by the university's IRB and was conducted under the guidance of the university's Fire Marshal.

C. Hypothesis

We hypothesize that in a situation where participants are currently experiencing risk and have experienced a robot's behavior in a prior interaction, participants will tend to follow guidance from an efficient robot but not follow guidance from a circuitous robot. Moreover, participant's self-reported trust will strongly correlate with their decision to follow or not follow the robot.

IV. Experimental Setup

All experiments took place in the office area of a storage building on our campus. The building was otherwise unoccupied during experiments. The office area contained a hallway and several rooms. The room at the end of the hallway was designated as the meeting room and the room next to it was designated as the other room, only used in the circuitous behavior condition. The back exit used for this experiment actually lead to a large storage area, but this was obscured using a curtain. Participants could see light through the curtain, but could not see the size of the room. This was intended to make this doorway into a plausible path to an exit, but not a definite exit to the outdoors. A standard green emergency exit sign hung in the hallway indicating that participants should exit through the main entrance in the event of an emergency. A room in the middle of the building was designated as the control room. An experimenter stayed in that room controlling

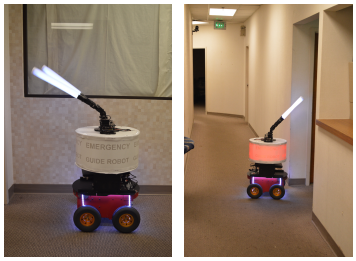

Figure 3. Robot during non-emergency phase of the experiment pointing to meeting room door (left) and robot during emergency pointing to back exit (right). Note that the sign is lit in the right picture. A standard emergency exit sign is visible behind the robot in the emergency.

the robot through an RF link. The experimenter could view the entire experiment area from five cameras placed throughout the building but could not be seen by participants.

The emergency guide robot (Figure 3) used a Pioneer P3-AT as a base. The base had white LED strip lights along all sides to illuminate the ground around it. A platform was built on top of this base to house a laptop computer and support a lighted cylindrical sign 24.5 cm tall and 47 cm in diameter. The words "EMERGENCY GUIDE ROBOT" in 3 cm tall letters were backlit by red LEDs. These LEDs were off during the non-emergency phase of the experiment but turned on during the emergency. Two PhantomX AX-12 Pincher arms were mounted to the top of the sign. Only the first three joints (the two shoulder servos and the elbow servo) on each arm were present. On top of each arm was a cylinder of foam lit with white LEDs. The arms, including foam, were 68 cm long. While the robot was moving the arms stayed straight up. The arms pointed straight ahead and oscillated by 20 degrees up and down to indicate that a participant should proceed in the direction the robot is facing (either into the meeting room or to the back exit). The robot measured 68 cm from ground to the top of the sign and 136 cm tall with arms fully extended up.

Artificial smoke was provided by a Bullex SG6000 smoke generator. The artificial smoke is non-toxic and non-carcinogenic. A First Alert smoke detector was placed on the hallway side of the doorframe of the meeting room door. Another of the same model was placed in the other room on the wall in case the first did not sound. The detectors alternated between producing a buzzing noise and the words "Evacuate! Smoke! Evacuate!" when they detected smoke. The alarm could easily be heard in the meeting room with the door closed.

Participants were recruited via emails to students at the university. Thirty participants were recruited for this study but four were not included in the results because they did not complete the experiment. Two participants did not leave the meeting room after the alarm sounded and had to be retrieved by experimenters. One participant activated the abort device after walking through the smoke and was intercepted by an experimenter before completing the experiment. In one trial, neither alarm sounded after the smoke filled the hallway, so the experiment was aborted. Of the 26 remaining participants (31% female, average age of 22.5), 13 were in each condition. All but three participants stated they were students. Participants were not warned that an emergency would occur.

Participants were advised before signing up for the experiment and in the consent form that they should not participate in this experiment if they had known heart conditions, asthma, other respiratory conditions, Post-Traumatic Stress Disorder (PTSD), anxiety disorders, or claustrophobia. They were not told why. These exclusion criteria were put in place because the artificial smoke can irritate people with respiratory disorders and because the emergency scenario could negatively affect participants with heart conditions or psychological disorders. Participants were also required to be between the ages of 18 and 50 (for health reasons) and capable of simple physical activity, such as walking around the building. The exclusion criteria was intentionally designed to be restrictive to ensure participant safety to the extent possible.

V. Results

The results from this experiment were surprising: all 26 participants followed the robot's instructions to proceed to the back exit in the emergency (Figure 4). Eighty-one percent of participants indicated that their decision to follow the robot meant they trusted the robot. The remaining five individuals (three in the efficient condition, two in the circuitous condition) stated that trust was not involved. They justified this with a variety of different reasons. One participant in the circuitous condition stated that they did not believe that the emergency was real. One in the efficient condition felt that they had no choice in the emergency. Another in the efficient condition noted that following the robot was the logical choice. One participant (also in the efficient condition) indicated that the robot was designed to help (and thus it was not the robot that was being trusted) and the last (in the circuitous condition) believed that trust was not involved in this interaction because they would not necessarily trust the robot in every emergency. Eighty-five percent of participants indicated that they would follow the robot in a future emergency. Only three participants noticed the emergency exit sign behind the robot and none expressed an interest in following it.

Results from the second survey found that just four of the thirteen participants with the circuitous robot reported that it was a bad guide. Three other participants indicated that it was a good guide in general, but that it made a mistake by going into the other room. The remaining six participants gave varying reasons for why they thought the robot was a good guide, including that it moved smoothly and pointed to the right room in the end. It is worth noting that in [10] the researchers found that many participants marked that the robot was a good guide in the non-emergency phase of the experiment, but were still biased against following it in the emergency. This result inspired one of the exploratory studies presented in the next section.

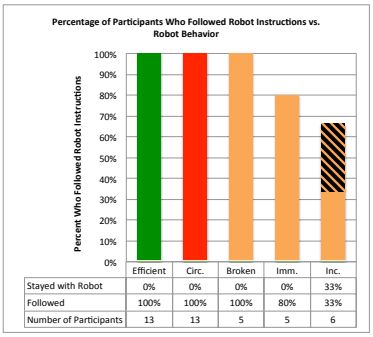

Figure 4. Results from the main study (green and red bars) and exploratory studies (orange bars) discussed in the next section.

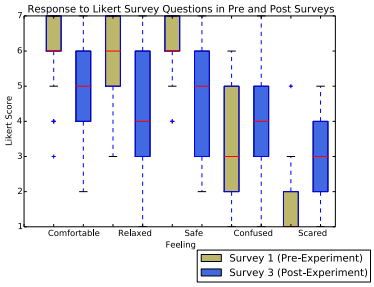

Figure 5. Change in participant responses to questions about their feelings from before the experiment (gold) to the emergency (blue).

There are confounding factors that could serve as alternative explanations for the results and explain why participants behaved differently in this experiment than in previous virtual emergency experiments such as [13], [10]. Lack of believability during the experiment is one confounding factor. Participants may not have believed the simulated emergency was real and based their decision and survey responses as such. It is difficult to measure the realism of the experiment because participants may not want to admit they believed it (social desirability bias). We attempted to evaluate the experiment's believability by asking participants to complete a survey about their current feelings before and after the experiment. The change in these results can be seen in Figure 5. All of the survey questions were on a 7-point Likert scale. Participants generally reported being comfortable, relaxed and safe before the experiment began (median of 6 for each). Some participants reported being confused (median of 3) and almost none reported being scared (median of 1) in the beginning. There was very little change (median changed less than or equal to 1 on each question) in the second survey. Participant answers in the third survey showed a marked change in comfort, relaxation, and safety level, (median of 5, 4, and 5, respectively), and increase in confusion (median of 4.5) with a similar increase for the scared scale (median of 2.5). Fifty-four percent of participants gave an increased confusion score between the pre and post surveys with 27% (seven participants) increasing that score by 3 or more. Additionally, 62% of participants (mainly the same participants) increased their response to the scared question with 15% increasing their rating by 3 or more. Wilcoxon Signed-Ranks Tests indicate that these results were significant: Comfortable $Z = 12, p < 0.001$, Relaxed $Z = 22, p = 0.003$, Safe $Z = 26, p < 0.001$, Confused $Z = 35.5, p = .023$, Scared $Z = 4.5, p < 0.001$

Despite this decrease in positive feelings and increase in negative feelings, most participants (58%) rated the realism of the emergency as low (a 1 or 2). Thirty-eight percent of participants rated it as moderate (3, 4 or 5) and only one participant rated it as high (a 6). The one participant who aborted the experiment (not included in the results above due to not completing the experiment) after seeing the smoke rated it a 6 out of 7. After reviewing video recordings of the experiment, we observed that 42% of participants had a clear physical response (either leaning away from the smoke or stepping back from the door in surprise) when opening the door to a smoke-filled hallway. This leads us to believe that many participants were likely exhibiting post-hoc rationalization on the survey: when they took the survey they knew that the experiment was not real, so they responded accordingly.

VI. Exploratory Studies: How to Bias Against Following the Robot

The results above are promising from one perspective: clearly the robot is considered to be trustworthy in an emergency. Yet, it is concerning that participants are so willing to follow a robot in a potentially dangerous situation even when it has recently made mistakes. The observed phenomena could be characterized as an example of overtrust [8]. Overtrust occurs when people accept too much risk believing that the trusted entity will mitigate this risk. This raises the important question, how defective must a robot be before participants will stop following its instructions?

Following our main study, we conducted three small studies to determine if additional behaviors from the robot either before or during the emergency would convince participants not to follow its instructions in the emergency. The first exploratory study, labeled Broken Robot, tested a new behavior during the non-emergency phase of the experiment. The second, Immobilized Robot, evaluated a behavior that spanned the entire study. The final study, Incorrect Guidance, determined the effect of odd robot behavior during the emergency phase of the experiment. A total of 19 participants were recruited for the three studies but three did not complete the experiment. One because the alarm failed to sound, one because the participant left the meeting room before the emergency started, and one because the participant did not leave the meeting room after the alarm sounded. The 16 remaining participants (38% female, average age of 20.9 years old) were divided into the three new conditions.

A. Broken Robot

We believed that the robot's behavior during the non-emergency phase of the experiment would influence the decision-making of the participant during the emergency.

Given that about half of the participants did not realize that the circuitous robot had done anything wrong, we designed a robot behavior that would obviously be a bad guide. As with the main experiment, this experiment began by guiding participants down the hallway. When it reached the first corner, the robot spun in place three times and pointed at the corner itself. No discernible guidance information was provided by the robot to participants. An experimenter then approached the participant and said, "Well, I think the robot is broken again. Please go into that room [accompanied with gestures to the meeting room] and follow the instructions. I'm sorry about that." The experiment then proceeded as in the previous conditions: the participant closed the meeting room door, the robot was moved to its emergency position (Figure 1) and smoke was released to begin the emergency phase. Five participants took part in this condition.

During the emergency, despite the robot's breakdown in the non-emergency phase of the experiment, all five participants followed the robot's guidance by exiting through the back exit (Figure 4). All five indicated that their decision meant that they trusted the robot and all five indicated that they would follow it in a future emergency. Four of the five participants indicated that the robot was not a good guide in the non-emergency phase of the experiment. The only one who indicated that it was a good guide did not hear the speech from the experimenter and thus did not experience the entire robot condition. The participant saw the robot spin in circles and then found the meeting room without any help. He considered that the robot had done a good job because he was able to find the meeting room quickly. Despite the higher percentage of participants who rated the robot as a bad guide in the non-emergency phase of the experiment, this condition produced the same results as in the circuitous condition.

Participants rated the emergency with a median of 3 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 5 for comfort, 5 for relaxation, 6 for safety, 4 for confusion and 4 for scared.

B. Immobilized Robot

In the immobilized robot condition, we attempted to convince participants that the robot was still malfunctioning during the emergency by having it behave poorly throughout the experiment.

At the start of the experiment, the robot moved a short distance forward, but then, upon reaching the intersection of the hallways (Robot Emergency Position in Figure 1) it spun in place three times and then pointed to the back exit. At this point, an experimenter informed the participant that the robot was broken with a similar speech as in the broken robot condition. The robot did not move and continued gesturing towards the back exit for the remainder of the experiment. The robot's lights were not turned on. From the perspective of an evacuating participant, the robot did not appear to have moved or changed behavior from when they were told it was broken in the non-emergency phase of the experiment. Five participants took part in this condition.

Figure 6. Robot performing incorrect guidance condition by pointing to a dark, blocked room in the emergency.

In this condition, four of the five participants followed the robot in the emergency (Figure 4). The one participant who did not follow the robot noticed the exit sign and chose to follow it instead. Three of the four participants who followed the robot's guidance indicated that they trusted it (the remaining said that this was the first exit available and thus trust was not involved). Two said they would follow it again in the future. All five rated the robot as a bad guide in the non-emergency phase of the experiment of the experiment.

Participants rated the emergency with a median of 1.5 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 3 for comfort, 3 for relaxation, 5 for safety, 6 for confusion and 4 for scared.

C. Incorrect Guidance

Inspired by the results in the immobilized robot condition, we tried a third robot behavior that might convince participants not to follow its guidance in an emergency. In this condition, the robot performed the same as in the broken robot condition, with accompanying experimenter speech, in the non-emergency phase of the experiment. During the emergency, the robot was stationed across the hall from its normal emergency position and instructed participants to enter a dark room (Figures 1 and 6). The doorway to the room was blocked in all conditions with a piece of furniture (initially a couch then a table when the couch became unavailable) that left a small amount of room on either side for a participant to squeeze through to enter the room. There was no indication of an exit from the participant's vantage point. All lights inside the room were turned off. Six participants took part in this condition.

Two of six participants followed the robot's guidance and squeezed past the couch into the dark room. An additional two participants stood with the robot and did not move to find any exit on their own during the emergency. Experimenters retrieved them after it became clear that they would not leave the robot. The remaining two participants proceeded to the front exit of the building (Figure 4). The two participants who followed the robot's instructions indicated that this meant they trusted the robot, although one said that he would not follow it again because it had failed twice. The two who stayed with the robot indicated that they did not trust the robot and the two who proceeded to the front exit selected that trust was not involved in their decision. None of those four indicated that they would follow the robot in a future interaction. All six participants wrote that the robot was a bad guide in the non-emergency phase of the experiment.

Participants rated the emergency with a median of 1.5 out of 7 on the realism scale. Participants rated their feelings in the emergency scenario with a median of 4 for comfort, 4 for relaxation, 5 for safety, 5.5 for confusion and 3 for scared.

VII. Discussion

Our results show that none of the robot behaviors performed solely in the non-emergency phase of the experiment had an effect on decisions made by participants during the emergency. These results conflict with our hypothesis and offer evidence that errors during prior interactions have little effect on a person's later decision to follow the robot's guidance. These results appear to disagree with the work of others examining operator-robot interaction in low-risk situations [3] and emergency guidance studies in virtual simulation environments [13], [10]. A similar conclusion was reached in [15]. We have found that participants have a tendency to follow a robot's guidance regardless of its prior behavior. To better understand participants' reasoning, we examined their survey response. Of the 42 participants included in all of our studies, 32 (76%) reported not noticing the exit sign behind the robot's emergency position. Upon turning the corner from the smoke filled hallway on their way out, participants' eyes were drawn to the large, well-lit, waving robot in the middle of their path. Couple the visual attraction of the robot with the increased confusion reported on the surveys and it is no surprise that participants latched onto the first and most obvious form of guidance that they observed.

These results are in contrast to previous results that found participants did not follow a previously bad robot in a virtual simulation of an emergency. In the high-risk scenario investigated here, participants observed what appeared to be smoke and had to make fast decisions. Although the virtual emergency was also under time pressure, participants were not in real danger and thus were able to be more deliberative in their decision-making. They were likely conscious of the fact that they were in no real danger and so they could take their time to make the best choice possible.

Several alternative explanations for the results are possible. Below, we give our opinions on these explanations, but more testing is necessary to conclusively eliminate them. One alternative explanation is that the age of the participants caused the observed results. Participants in this study were mostly university students and therefore younger and possibly more accepting of new technology than a more diverse population. Still, even if our findings are only true in relation to a narrow population, they show a potentially dangerous level of overtrust.

The realism of the scenario is addressed in detail above, but still presents an alternative explanation. Perhaps participants did not believe that they were in any danger and followed the robot for other reasons. Their increased confusion scores and reactions to the smoke indicate that at least some of the participants were reacting as if this was a real emergency. Given that every participant in the initial study followed the robot, regardless of their reaction to the emergency, we conclude that the realism of the scenario had little or no effect on their response. Additionally, many participants wrote that they followed the robot specifically because it stated it was an emergency guide robot on its sign. They believed that it had been programmed to help in this emergency. This is concerning because participants seem willing to believe in the stated purpose of the robot even after they have been shown that the robot makes mistakes in a related task. One of the two participants who followed the robot's guidance into the dark room even thought that the robot was trying to guide him to a safe place after he was told by the experimenter that the exit was in another direction. It is possible that participants saw the robot as an authority figure; however, this leads to further questions about why participants would trust such an authority figure after it had already made a mistake.

It is worth mentioning that many people in real-life fire drills and fire emergencies do not believe that they are in real danger (see [4] for an example using the 1993 World Trade Center bombing). Some participants wrote on their surveys that the fire alarm used in this experiment sounded fake, even though it was an off-the-shelf First Alert smoke detector. Others stated that the smoke seemed fake, even though this same artificial smoke is used to train firefighters. It is likely that participants would respond the same when encountering real smoke.

Perhaps participants only followed the robot because they felt that they should do so in order to complete the experiment. In fact, researchers have found that participants were more likely to respond positively to automation that displayed good etiquette, so it is possible that participants were only following the robot to be polite [9]. One participant of the 42 tested wrote that he followed the robot only because he was told to in the non-emergency phase of the experiment. Each of the conditions in the exploratory studies attempted to realign participant beliefs by having the experimenter interrupt the robot and lead the participant himself. In the broken and immobilized robot case, nine of ten participants still followed the robot in the emergency. Thus, we do not believe that etiquette or prior instructions explain our results.

A final alternative explanation is that the building layout was sufficiently simple that participants believed that they had ample time to explore where the robot was pointing and still find their way out without being harmed. This is possible, but participants did not express a desire to explore any other rooms or hallways in the building, just the one pointed to by the robot. Some participants looked into the other room on their way out, but none spent time exploring it. No participant tried to open either of the closed doors on their way out and, except in the incorrect guidance case, no participant tried to enter either of the rooms blocked by furniture. Participant behavior appears to reflect their conviction to follow the robot's guidance and their survey responses indicate that they believed the robot was guiding them to an exit.

VIII. Conclusion and Future Work

Prior to conducting the experiment, we expected that participants would need to be convinced to follow a robot in

an emergency, even if they did not believe the emergency was real. It is reasonable to assume that a new technology is imperfect, so new life-saving (and therefore life-risking) technology should be treated with great caution. Informal discussions with several prominent roboticists and search-and-rescue researchers reinforced this idea. In contrast, we found that participants were all too willing to trust an emergency guide robot, even when they had observed it malfunction before. The only method we found to convince participants not to follow the robot in the emergency was to have the robot perform errors during the emergency. Even then, between 33% and 80% of participants followed its guidance.

This overtrust gives preliminary evidence that robots interacting with humans in dangerous situations must either work perfectly at all times and in all situations or clearly indicate when they are malfunctioning. Both options seem daunting. Our results indicate that one cannot assume that the people interacting with a robot will evaluate the robot's behavior and make decisions accordingly. Additionally, our participants were willing to forgive or ignore robot malfunctions in a prior interaction minutes after they occurred. This is in contrast to research on operator-robot interaction, which has shown that people depending on a robot are not willing to forgive or forget quickly.

These results have important ramifications for the study of human-robot interaction. The results highlight the impact of the environment on the decision-making of a person in regard to a robot, although more research is needed before firm conclusions are drawn. In high-risk situations people may blindly follow or accept orders from a robot without much regard to the content or reasonableness of those instructions. It may be hard for the robot to cede control back to the person in these situations.

This study also opens many directions for future work. The most obvious direction is to understand the factors that contribute to overtrust. For instance, discerning if certain personality types or defining which types of situations increase one's susceptibility to overtrust is an important next step. Developing techniques to prevent overtrust is another important direction for future work. Ideally, these techniques would allow a person to calibrate their trust in a system, engendering an appropriate level of trust fitted for the robot's capabilities. Many additional questions are raised by our results. How does a robot inform nearby people that it is malfunctioning and should not be trusted? Will frightened evacuees listen to the robot when it tells them to stop following it and find their own way out? Can a non-verbal robot communicate such a message with its motion alone? How many errors must a robot make before it loses an evacuee's trust?

Human-robot trust has become a very important topic as autonomous robots take on more tasks in the real world. Self-driving cars and package delivery drones represent a much greater risk to people than floor-cleaning robots. We must understand the factors that affect trust in these autonomous systems. Additionally, we must understand that people might overtrust a robot to perform its given task, regardless of the robot's prior performance and find ways to mitigate the risk that overtrust brings.

IX. Acknowledgements

The researchers would like to thank Larry Labbe and the Georgia Tech Fire Safety Office for their support during this research. Support for this research was provided by the Linda J. and Mark C. Smith Chair in Bioengineering, Air Force Office of Sponsored Research contract FA9550-13-1-0169 and Georgia Tech Research Institute.

References

[1] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1):41–52, 2011.

[2] L. Benthorn and H. Frantzich. Fire alarm in a public building: How do people evaluate information and choose an evacuation exit? *Fire and Materials*, 23(1):311–315, 1999.

[3] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 251–258. IEEE Press, 2013.

[4] Rita F Fahy and Guylene Proulx. Human behavior in the world trade center evacuation. *Fire Safety Science*, 5:713–724, 1997.

[5] G. Gao, A. A. Clare, J. C. Macbeth, and M. L. Cummings. Modeling the impact of operator trust on performance in multiple robot control. In *2013 AAAI Spring Symposium Series*, 2013.

[6] W. Grosshandler, N. Bryner, D. Madrzykowski, and K. Kuntz. Report of the technical investigation of The Station Nightclub Fire. Technical report, National Institute of Standards and Technology, 2005.

[7] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527, 2011.

[8] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.

[9] Raja Parasuraman and Christopher A Miller. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4):51–55, 2004.

[10] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. Timing is key for robot trust repair. In *Seventh International Conference on Social Robotics*, 2015.

[11] Paul Robinette, Patricio A. Vela, and Ayanna M. Howard. Information propagation applied to robot-assisted evacuation. In *2012 IEEE International Conference on Robotics and Automation*, 2012.

[12] Paul Robinette, Alan R. Wagner, and Ayanna M. Howard. Assessment of robot guidance modalities conveying instructions to humans in emergency situations. In *RO-MAN*. IEEE, 2014.

[13] Paul Robinette, Alan R. Wagner, and Ayanna M. Howard. The effect of robot performance on human-robot trust in time-critical situations. Technical Report GT-IRIM-HumAns-2015-001, Georgia Institute of Technology. Institute for Robotics and Intelligent Machines, Jan 2015.

[14] Daniel Safarik and Antony Wood. An all-time record 97 buildings of 200 meters or higher completed in 2014. In *CTBUH Year in Review*, 2014.

[15] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–148. ACM, 2015.

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

[17] Alan R. Wagner and Paul Robinette. Towards robots that trust: Human subject validation of the situational conditions for trust. *Interaction studies*, 16(1), 2015.