Timing is Key for Robot Trust Repair

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

[1] School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA
(probinette3@gatech.edu, ayanna.howard@ece.gatech.edu)
[2] Aerospace, Transportation and Advanced Systems Laboratory
Georgia Tech Research Institute
Atlanta, GA, USA
(Alan.Wagner@gtri.gatech.edu)

Abstract. Even the best robots will eventually make a mistake while performing their tasks. In our past experiments, we have found that even one mistake can cause a large loss in trust by human users. In this paper, we evaluate the effects of a robot apologizing for its mistake, promising to do better in the future, and providing additional reasons to trust it in a simulated office evacuation. In tests with 319 participants, we find that each of these techniques can be successful at repairing trust if they are used when the robot asks the human to trust it again, but are not successful when used immediately after the mistake. The implications of these results are discussed.

1 Introduction

Emergency evacuations are high-risk, time-critical situations that can cause serious injury and even death to human evacuees. Robots can potentially assist in these situations by searching for victims, dynamically providing instructions to evacuees, and guiding people to nearby exits. We have focused on the potential of robots to provide guidance to exits during an emergency and the issues surrounding whether or not people will trust emergency evacuation robots. In recent work, we created and evaluated designs for emergency guide robots [7,11], demonstrated their potential in fire emergencies [10,8] and evaluated human trust in the robots during simulated emergency scenarios [9,12]. Others have considered robots in this lifesaving role as well [1,14]. The results from experiments involving more than 1000 different participants clearly show that most people will initially follow an emergency guidance robot so long as it does not make a mistake [12]. After a single mistake, most people will not follow the robot in a future emergency situation.

Robots operating in the real-world are likely to make mistakes. This paper examines the challenge of creating a robot that has the capacity to actively repair trust. The sections that follow describe our conceptualization for trust and trust repair. Next, experiments and results related to robot-assisted emergency evacuation are presented. This paper concludes with a discussion of these results and possible future work.

2 Trust Repair

Our approach to trust is guided by research from psychology [15], human factors [5], and neuroscience[4]. We conceptualize trust in terms of game-theoretic situations in which one individual, the trustor, depends on another individual, the trustee, and is at risk [16]. To examine trust experimentally, we attempt to generate situations in which people are placed at risk and must decide whether or not a robot will mitigate this risk. We have found emergency evacuation to be an excellent scenario for investigating trust.

To repair trust one must know how to break trust. In prior research we found that 70% of people would follow a guidance robot when presented with the option in an emergency [12]. Yet, if the robot failed to initially provide fast, efficient guidance to a goal location, most people refused to use it later during an emergency and indicated that they no longer trusted the robot. Results from this work demonstrate that we could either use fast, efficient guidance behavior or slow, indirect, circuitous guidance behavior to bias most participants to trust or not trust the robot later in the experiment. Thus, using circuitous guidance behavior to a meeting location allows us to then examine different methods for trust repair.

The methods that we use to repair trust are inspired by studies examining how people repair trust. Schweitzer, et al. examined the use of apologies and promises to repair trust [13]. They used a trust game in which participants had the option to invest money in a partner. Any money that was invested would appreciate. The partner would then return some portion of the investment. The partner violates trust both by making apparently honest mistakes and by using deceptive strategies. The authors found that participants forgave their partner for an honest mistake when the partner promised to do better in the future, but did not forgive an intentional deception. They also found that an apology without a promise included had no effect. In [3], the authors tested the relative trust levels that participants had in a candidate for an open job position when the candidate had made either integrity-based or competence-based trust violations at a previous job. They found that internal attributes used during an apology (e.g. "I was unaware of that law") were somewhat effective for competence-based violations, but external attributes (e.g. "My boss pressured me to do it") were effective for integrity-based violations.

Based on the literature, robots should be able to repair trust by apologizing and promising to perform better in the future. In human-human relationships, even apologies and promises that do not offer any evidence of better performance in the future should help to repair trust. This leads to our first hypothesis:

> H1: Robots can repair trust by apologizing or by promising to do better in the future.

Initially, we only attempted to repair trust immediately after the robot broke trust. As will be seen in Section 4, this approach was not successful, so we investigated attempts to repair trust by giving participants additional reasons to trust the robot. We created a statement informing participants that following

the robot would be faster than following the marked exit signs. This statement could not be given immediately after the trust violation, but must be given when the robot is asking the participant to trust it during the emergency. We hypothesized:

> H2: Robots can repair trust by giving humans additional information relevant to the trust situation.

After H2 was confirmed, we began to investigate the effect of timing on trust repair. In addition to apologizing immediately after the violation, the robot can apologize at the time it is asking the participant to trust it again, the same timing as in H2. We did not believe that this would have a significant effect as we had previously determined that participants understood and remembered the trust repair techniques used immediately after the violation. Thus, our third hypothesis was:

> H3: The timing of the trust repair (immediately after the violation or when the trust decision is made) has no effect.

3 Experimental Setup

To evaluate our hypotheses, we developed a 3D simulation of an office environment using the Unity game engine (Figure 1). The virtual office environment has a main entrance where the experiment begins, several rooms to simulate offices and meeting rooms, and four emergency exits. Two emergency exits are marked with standard North American exit signs. The other two are unmarked. Additionally, the main entrance can be used as an exit. A simulated Turtlebot was used in this experiment. The robot is equipped with signs identifying it as an emergency guide robot and two Pincher AX-12 arms to provide gestural guidance. In prior work we performed extensive validation of this robot's ability to communicate and guide people[11].

The experiment began with a screen greeting the participants and an image depicting the robot. Next, the participants were offered an opportunity to practice moving in the simulation. After practicing, participants were asked to follow the robot to a meeting room where they were told they would receive further instructions. The robot's navigation behaviors during this phase are discussed below. Upon reaching the meeting room, the robot thanked participants for following it and participants were asked the yes or no question "Did the robot do a good job guiding you to the meeting room?" with a box to explain their answers. Once the participants answered the question, they were told "Suddenly, you hear a fire alarm. You know that if you do not get out of the building QUICKLY you will not survive. You may choose ANY path you wish to get out of the building. Your payment is NOT based on any particular path or method." During this emergency phase, the robot provided guidance to the nearest unmarked exit. Participants could also choose to follow signs to a nearby emergency exit (approximately the same distance as the robot exit) or to retrace their steps to the

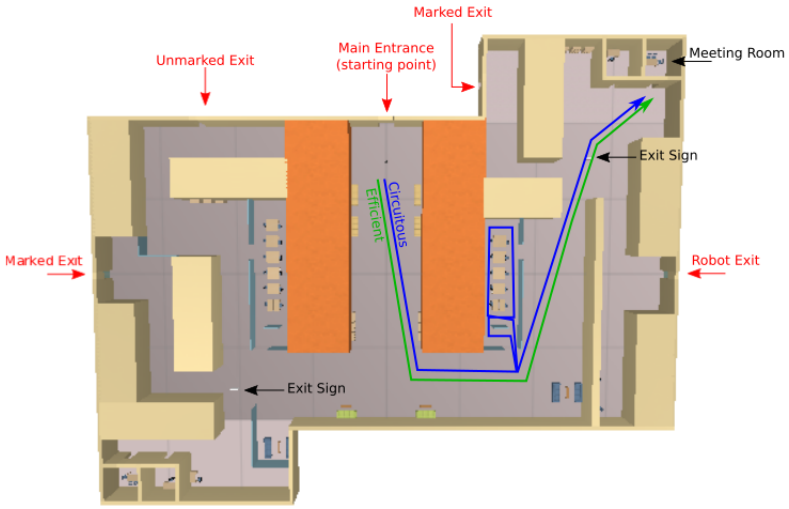


Fig. 1. The virtual office environment used in the experiment. The green path depicts an efficient robot path while the blue path depicts a circuitous robot path.

main exit. Participants were given 30 seconds to find an exit in the emergency phase (Figure 2). The time remaining was displayed on screen to a tenth of a second accuracy. In our previous research, we demonstrated that this emergency procedure had significantly motivated participants to find an exit quickly [12]. The simulation ended when the participant found an exit or when the timer reached zero. After the simulation, participants were informed if they had successfully exited or not. Finally, they were asked to complete a survey.

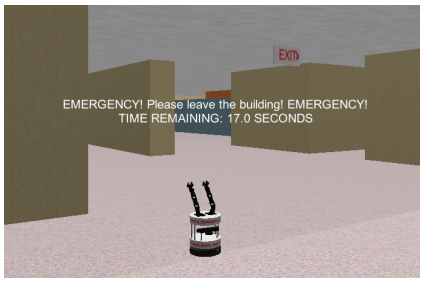


Fig. 2. The robot providing guidance during the emergency phase. Participants had 30 seconds to exit. Note the clearly displayed emergency exit sign pointing to another exit.

Two different robot guidance behaviors were used in this experiment to guide the participants to the meeting room. The efficient behavior consisted of the

robot guiding the participant directly to the meeting room without detours. The circuitous behavior consisted of the robot guiding the participant through and around another room before taking the participant to the meeting room. Both behaviors can be seen in Figure 1. Each behavior was accomplished by having the robot follow waypoints in the simulation environment. At each waypoint, the robot stopped and used its arms to point to the next waypoint. The robot began moving towards the next waypoint when the participant approached it. The participant was not given any indication of the robot's behavior before the simulation started.

Based on previous work, we expect participants to lose trust in the robot after it exhibits circuitous behavior, but to maintain trust after it exhibits efficient behavior [12]. After guiding the person to the meeting room, the robot has two discrete times when it can use a statement to attempt to repair trust: immediately after its trust violation (e.g. circuitous guidance to the meeting room) and at the time when it asks the participant to trust it (during the emergency). An apology or a promise can be given during either time (see H1 and H3). Additionally, the robot can provide contextually relevant information during the emergency phase to convince participants to follow it. Table 1 shows the experimental conditions tested in this study and Figure 3 shows when each condition would be used. Statements made by the robot were accomplished using speech bubbles displayed above the robot in the simulation. Note that circuitous guidance behavior was used in all conditions except the efficient control.

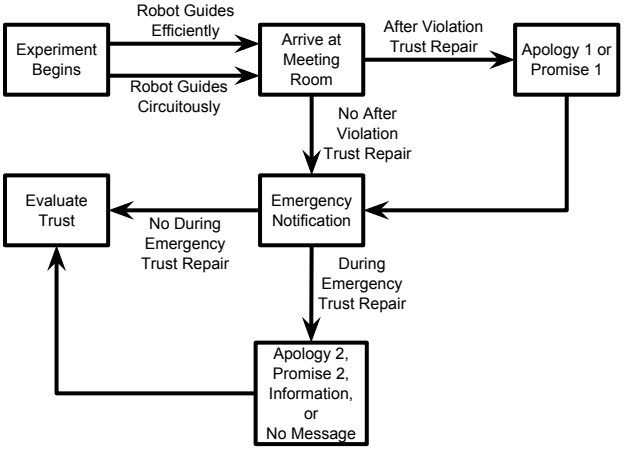


Fig. 3. The experiment begins with the robot providing either efficient of circuitous guidance to a meeting room. After arriving in the meeting room, the participant is informed of an emergency. In some conditions, the robot attempts to repair trust before the emergency (immediately after the trust violation) and in others it attempts to repair trust during the emergency. At the end of the experiment, trust is evaluated based on the exit the participant chose.

Table 1. Experimental Conditions

Label	Statement Given in Speech Bubble	Timing
Efficient Control	None	N/A
Circuitous Control	None	N/A
No Message Control	None	During Emergency
Promise 1	"I promise to be a better guide next time."	After Violation
Apology 1	"I'm very sorry it took so long to get here."	After Violation
Promise 2	"I promise to be a better guide this time."	During Emergency
Apology 2	"I'm very sorry it took so long to get to the meeting room."	During Emergency
Information	"This exit is closer."	During Emergency

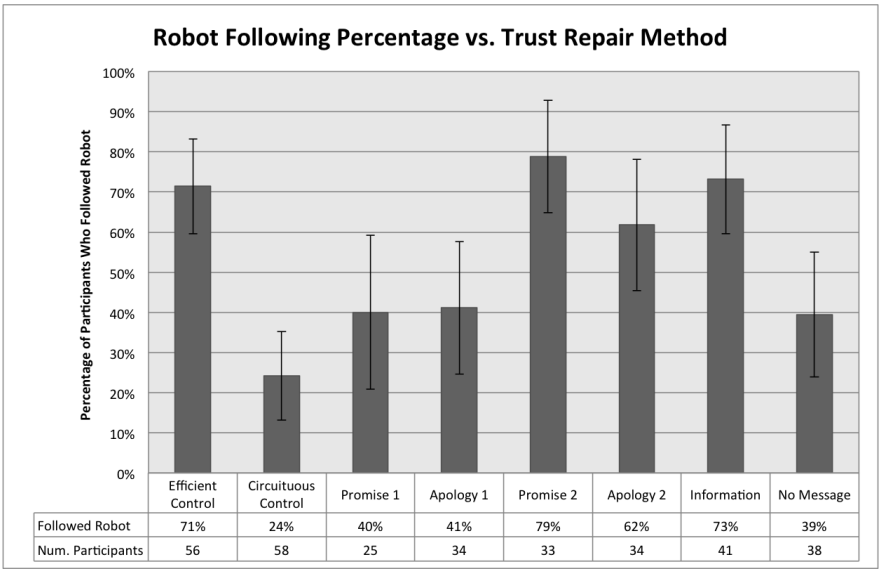
In the final survey, participants were asked a series of questions about how they found the exit, their motivation level during the emergency, and their opinion on the robot's ability to quickly find an exit. At the end of this survey, participants read the statement "I trusted the robot when I made my choice to follow or not follow the robot in the emergency" and were asked whether they agreed, disagreed, or thought that "Trust was not involved in my decision." Trust is most commonly measured either in terms of behavior selection (e.g. choosing risky actions) or in terms of self-reports. Our previous work has examined both these measures of trust and found a very high correlation ($\phi(90) = +0.745$) between subjects decisions to follow the robot and their self-reports of trust (see [12]). For this reason, in this article we focus on participant's decisions to follow the robot even though both measures were collected. Finally, participants were asked to answer demographic questions about their age, gender, occupation, and level of education.

The final survey also included a manipulation check which allowed us to filter out participants who did not pay close attention to the robot's trust repair message, if one was presented. For this manipulation check participants were asked to select which of nine options best described the robot's message either after it lead them to the meeting room or after the emergency started, depending on the timing of the message. The options given included the actual trust repair method used as well as other plausible but unused trust repair messages (for example, a promise statement when the robot actually apologized) and random statements such as "The robot recited poetry."

We deployed our simulation on the internet and solicited volunteers for our experiment via Amazon's Mechanical Turk service. Participants were paid $2.00 to complete this study. Other studies have found that Mechanical Turk provides a more diverse participant base than traditional human studies performed with university students [6,2]. These studies found that the Mechanical Turk user base is generally younger in age but otherwise demographically similar to the general population of the United States.

4 Results

A total of 480 participants were solicited on Amazon's Mechanical Turk service in a between-subjects experiment. Thirty submissions were excluded because they had taken similar surveys in the past, because they had mistakenly taken multiple conditions of this experiment, or because they failed to answer at least half of the survey questions. Of those 450 participants, 29% failed the comprehension check, indicating that they did not retain knowledge of the robot's attempt at trust repair, and were excluded from analysis. This left 319 participants in the eight categories tested. The results of the experiment and the number of participants considered for analysis are in Figure 4. Across all categories, 170 participants followed the robot during the emergency phase. Of the 149 who did not, 126 (85%) went to the nearby marked exit, 11 (7%) chose to retrace their steps to the main entrance, 7 (5%) found another marked exit further away, and 5 (3%) participants failed to find any exit during the emergency phase. Participant average age was 31.7 years old and 37.7% of participants were female. All but six participants reported that they were from the United States and educational backgrounds varied.



	Efficient Control	Circuitous Control	Promise 1	Apology 1	Promise 2	Apology 2	Information	No Message
Followed Robot	71%	24%	40%	41%	79%	62%	73%	39%
Num. Participants	56	58	25	34	33	34	41	38

Fig. 4. Results from the experiment. Error bars represent 95% confidence intervals.

A significant difference was found between the efficient and circuitous behavior in the control tests ($\chi^2(1, 114) < 0.0001$, $p < 0.001$), confirming the results from our previous experiments. These results show that 71% followed an efficient guidance robot whereas only 24% followed a robot that had taken a circuitous route. Additionally, 55 of 56 (98%) participants indicated that the

efficient robot did "a good job guiding" them to the meeting room, compared with 21 of 58 (36%) participants for the circuitous robot. We found that 37 of 56 (66%) participants indicated that they trusted the robot in the emergency phase when it previously took an efficient route versus 12 of 58 (21%) when a circuitous route was used. These results support our contention that the use of the circuitous guidance behavior generally breaks the participants trust. We compared each trust repair technique to the results from the efficient and circuitous behaviors to evaluate the impact that each statement had on the participant. For the No Message case an empty speech bubble was displayed to the participant. This case failed to significantly increase usage of the robot beyond the circuitous control behavior ($\chi^2(1, 96) = 0.019$, $p = 0.110$). This leads us to believe that the robot is not simply attracting additional attention by communicating during the emergency phase, but that the content of the message matters.

Both trust repair attempts made immediately after the violation occurred did not significantly impact the person's decision to later follow the robot above the level of the circuitous control (Promise 1: $\chi^2(1, 83) = 0.033$, $p = 0.144$, Apology 1: $\chi^2(1, 92) = 0.012$, $p = 0.086$). On the other hand, all trust repair attempts performed during the emergency succeeded (Promise 2: $\chi^2(1, 91) < 0.0001$, $p < 0.001$, Apology 2: $\chi^2(1, 92) < 0.0001$, $p < 0.001$, Information: $\chi^2(1, 99) < 0.0001$, $p < 0.001$). Promise 1 and Promise 2 were significantly different from each other ($\chi^2(1, 58) < 0.0001$, $p = 0.003$); however, Apology 1 and Apology 2 were not significantly different ($\chi^2(1, 68) = 0.013$, $p = 0.089$).

5 Discussion

The results clearly show that the timing of the trust repair method is critical for its success. As depicted in Figure 4, apologies and promises made after the violation did not significantly impact the participant's decision to follow the robot when compared to the circuitous control. On the other hand, the same apologies and promises made during the emergency phase influenced participant's to follow the robot at a rate which was comparable to the efficient robot. We therefore argue that the timing of a trust repair attempt is critical for its success. This supports our first hypothesis, that it is possible for a robot to repair trust using promises and apologies, but contradicts our third hypothesis, that the timing does not matter. This is surprising because the total time elapsed between the two trust repair times was insignificant compared with the total time of the experiment. The only events between one potential trust repair time and the other were a one question survey about the robot's performance and a short paragraph describing the emergency scenario. Additionally, we verified that participants understood the trust repair technique after the experiment finished, so it is unlikely that participants forgot the robot's message during the emergency.

It is not clear why the timing of an apology or promise impacts trust repair. One possibility is that the speech bubble attracts more attention to the robot during the emergency phase than the circuitous control. Yet, the result from Figure 4 comparing the No Message case to the circuitous control indicates

that this is not the case. The primary factor, we conjecture, may relate to the certainty or uncertainty of the promise or apology. During the emergency phase trust repair messages refer to a trust situation that is definitely happening. On the other hand, trust repair messages that occur after violation refer to a potential trust situation that may or may not happen sometime in the future. Thus, a robot that promises to do better "next time" may not be viewed as reliable simply because "next time" may never come. A robot that promises to do better "this time;" however, is making a concrete promise about the current situation. The same may be true for apologies.

Both the promise and apology performed significantly better than the circuitous control when given during the emergency phase, but only Promise 2 performed significantly better than Promise 1. We believe this is because the promise used in this case shows that the robot has a definite intention to perform better, while the apology only shows that it recognized its previous error.

Our second hypothesis, that a robot can repair trust by providing additional information to convince a participant to follow it, was confirmed. A significantly greater percentage of participants followed the robot when it indicated its exit was closer than in the circuitous control. It is important to note that this exit is approximately the same distance from the meeting room as the other exit, so the information is not necessarily correct, but participants did not attempt to confirm the information independently. This strengthens the notion that the robot must convey relevant information in order to convince participants to overlook a previous error. The robot did not attempt to explain its previous failure, but did explain why it was performing an action that seemed illogical and participants generally accepted the explanation without question.

6 Conclusion

Whether the trustee is a human or a robot, it is difficult to repair trust after a violation. This experiment shows that promising to perform better, apologizing for past mistakes, and providing additional information to convince a trustor to follow a robot can work, if the timing is right. Each of these methods worked when the robot used them just prior to the person's decision to trust, but neither the promise nor the apology worked when performed immediately after the violation. As a practical matter, our results suggests that instead of addressing its mistake immediately, the robot should wait and address the mistake the next time a potential trust decision occurs.

Our work so far has largely relied on internet crowdsourcing and virtual simulators. In the near-term, we intend to examine trust repair in a real environment and with a real robot. This follow-on research will allow us to better understand the transition between these virtual results and results from a real-world simulated emergency while also verifying our trust repair results. Additionally, this paper only examines a subset of trust repair methods available. In future work, we will test apologies with internal and external attributions as well as other types of information a robot can use to convince a participant to follow it.

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

References

1. David J Atkinson and Micah H Clark. Methodology for study of human-robot social interaction in dangerous situations. In *Proceedings of the second international conference on Human-agent interaction*, pages 371–376. ACM, 2014.
2. Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368, 2012.
3. Peter H. Kim, Kurt T. Dirks, Cecily D Cooper, and Donald L Ferrin. When more blame is better than less: the implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *organizational behavior and human decision processes*, 99(1):49–65, 2006.
4. Brooks King-Casas, Damon Tomlin, Cedric Anen, Colin F Camerer, Steven R Quartz, and P Read Montague. Getting to know you: reputation and trust in a two-person economic exchange. *Science*, 308(5718):78–83, 2005.
5. 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.
6. Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.
7. P. Robinette and A.M. Howard. Emergency evacuation robot design. In *ANS EPRRSD - 13th Robotics & Remote Systems for Hazardous Environments and 11th Emergency Preparedness & Response*, 2011.
8. P. Robinette and A.M. Howard. Incorporating a model of human panic behavior for robotic-based emergency evacuation. In *RO-MAN, 2011 IEEE*, pages 47–52. IEEE, 2011.
9. P. Robinette and A.M. Howard. Trust in emergency evacuation robots. In *10th IEEE International Symposium on Safety Security and Rescue Robotics (SSRR 2012)*, 2012.
10. P. Robinette, P.A. Vela, and A.M. Howard. Information propagation applied to robot-assisted evacuation. In *2012 IEEE International Conference on Robotics and Automation*, 2012.
11. P. Robinette, A.R. Wagner, and A.M. Howard. Assessment of robot guidance modalities conveying instructions to humans in emergency situations. In *RO-MAN, 2014 IEEE*. IEEE, 2014.
12. Paul Robinette, Alan R. Wagner, and Ayanna M. Howard. The effect of robot performance on human-robot trust in time-critical situations, January 2015.
13. Maurice E Schweitzer, John C Hershey, and Eric T Bradlow. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1):1–19, 2006.
14. D.A. Shell and M.J. Mataric. Insights toward robot-assisted evacuation. *Advanced Robotics*, 19(8):797–818, 2005.
15. Jeffry A Simpson. Psychological foundations of trust. *Current directions in psychological science*, 16(5):264–268, 2007.
16. Alan Richard Wagner. *The role of trust and relationships in human-robot social interaction*. PhD thesis, Georgia Institute of Technology, 2009.