When Should a Robot Apologize? Understanding How Timing Affects Human-Robot Trust Repair

Mollik Nayyar(✉) and Alan R. Wagner

The Pennsylvania State University, University Park, PA 16802, USA
{mxn244, alan.r.wagner}@psu.edu

Abstract. If robots are to occupy a space in the human social sphere, then the importance of trust naturally extends to human-robot interactions. Past research has examined human-robot interaction from a number of perspectives, ranging from overtrust in human robot interactions to trust repair. Studies by [15] have suggested a relationship between the success of a trust repair method and the time at which it is employed. Additionally, studies have shown a potentially dangerous tendency in humans to trust robotic systems beyond their operational capacity. It therefore becomes essential to explore the factors that affect trust in greater depth. The study presented in this paper is aimed at building upon previous work to gain insight into the reasons behind the success of trust repair methods and their relation to timing. Our results show that the delayed trust repair is more effective than the early case, which is consistent with the previous results. In the absence of an emergency, the participant's decision were similar to those of a random selection. Additionally, there seem to be a strong influence of attention on the participants' decision to follow the robot.

Keywords: Social robotics · HRI · Trust · Trust repair

1 Introduction

Trust in human interpersonal interactions is an integral component of human social behavior. It facilitates a number of fundamental interactions that are essential for our economic and social systems. Robots will play an increasingly important role in the human social sphere in the near future. It is therefore valuable to examine the concept of trust for human-robot interactions.

A variety of applications are currently being explored for robots to assist human beings in everyday life. One such application is robot assisted emergency evacuation [18, 19]. Robot assisted emergency evacuation may save lives by being constantly vigilant and providing valuable situation awareness to first responders. However, since the reliability of robots cannot be guaranteed, trusting these systems can potentially put evacuees and first responders at risk. Since robots will be used in multiple domains such as transportation, healthcare, and the military, developing an understanding of human-robot trust is crucial for safe introduction of robotic applications.

S. S. Ge et al. (Eds.): ICSR 2018, LNAI 11357, pp. 265–274, 2018.
https://doi.org/10.1007/978-3-030-05204-1_26

Past research in this domain has highlighted various aspects of human robot trust. In particular, [14] show that during emergencies, even in cases where the robot exhibits poor prior performance, people nevertheless tend to rely on it rather than their own instincts. In cases where trust is violated due to poor performance, it has been shown that a robot can repair trust by promising to do better or apologizing for mistakes if the robot promises or apologies at the right time [15]. The study in this paper is aimed at building upon this previous work to tease apart why the timing of trust repair statements impact a person's trust in an autonomous system. We hope to identify factors that affect trust repair. Any insight gained here will help us develop better models of trust from a human-robot perspective and will aid in our understanding of trust in general.

The following sections first present a small portion of the human-robot trust literature, focusing primarily on research related to trust repair. Next we present insights related to the how timing may impact trust repair. We then introduce our experimental setup and the different experimental cases are discussed. We conclude with results from simulation experiments involving 558 human subjects and a discussion of the insights these findings offer towards understanding human-robot trust repair.

2 Related Work

Researchers generally agree that trust-based decisions are characterized by situation in which the trustor is vulnerable and/or at risk and the actions of another individual can relieve or mitigate that risk [6, 8]. For humans interacting with robots or automated systems, human-like features such as politeness, facial features, and the system's speech, have been shown to increase trust [10, 12]. Humans also show a tendency to initially trust automated systems [2, 3, 7], even when they have no experience with the system. With experience, a person's trust in a system may be based on performance [7]. Handcock et al. [4] found that robot's performance had the strongest effect on trust. Yet it has also been shown that people will quickly come to overtrust automated systems and robots [11].

Most of these results have utilized relatively low-risk experimental paradigms such as economic games [5] or use of automated avatars for automated decision-making [13]. In contrast, our research focuses on human-robot trust in physically risky situations such as during search and rescue scenarios. Comparatively few studies have explored trust and the use of robots in emergency scenarios. Atkinson and Clark looked at different methodologies of studying human-robot interaction in a dangerous situation and found that human behavior carries forward to virtual environments and virtual simulations can be an effective method to study human-robot interactions [1].

Our own work has examined a variety of aspects related to robot led emergency evacuation. We have explored robot appearance [16], communication techniques [17], overtrust on a physical robot [14], and trust repair [15]. The research presented here builds upon our previous work on trust repair during emergency evacuations which shows that mistakes by a robot result in a sharp decrease in trust after the mistake, yet it was also shown that trust can be repaired if the robot apologizes or promises to do better. *Most importantly, this work demonstrated that the effectiveness of a trust repair* *statement strongly depends on the timing of the statement's delivery*. This paper attempts to dissect the reasons why the timing of a trust repair message by a robot impacts a person's trust in the system.

3 Trust Repair Timing

In our prior simulation research, a robot offered to lead a human subject to a meeting room, but made mistakes enroute to the room, eventually arriving at the location. In one condition, the robot apologized or promised immediately after the mistake. An emergency then occurred while the subject was in the meeting room and the subject was informed that they needed to quickly find an exit or their character would perish. The same robot offered to lead the subject to an exit. In a second, different condition, the robot apologized just prior to the subject deciding whether or not to follow the robot during the emergency. Figure 1 depicts the timing of the study's stages.

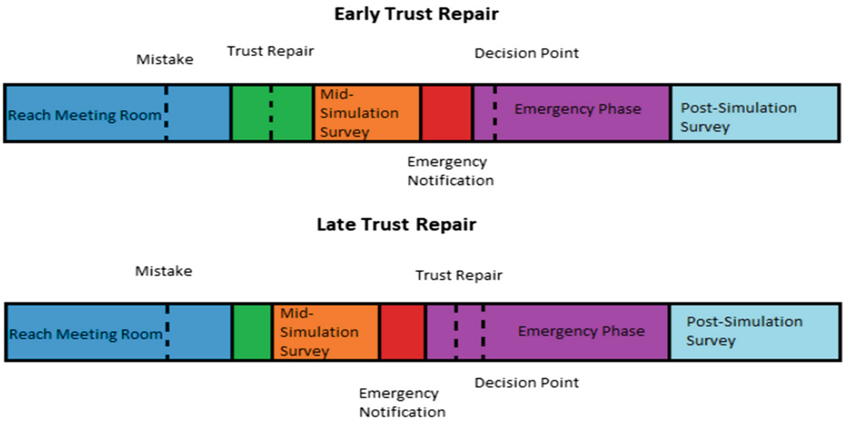


Fig. 1. A timeline of events in the experiment is depicted. The key difference is when the mistake, trust repair and decision point occurred. (Color figure online)

This research showed that only 40% of subjects followed the robot when it apologized or promised just after a mistake yet, 79% choose to follow the robot if it made the same apology or promise during the emergency [15]. Non-repair messages such as greetings or otherwise innocuous statements, on the other hand, do not repair trust. The study also revealed that trust does break when the robot makes a mistake and that the emergency strongly motivates people to find an exit quickly. But it was unclear why the timing of the trust repair statement (apology or promise) had such a large effect on people's decision whether or not to follow the robot.

The research presented here investigates several different hypotheses as to why the timing of trust repair statement might impact a person's decision to trust. Our previous experiments suggest several potential factors that might influence the importance of timing of trust repair messages. Our first step was to reproduce our prior results. Next, we examined if it is possible that the trust repair does not need to be read and internalized in order to influence the person. It might be possible that simply presenting a message will subconsciously influence the person. If this is the case, then we predict that reducing the display time of the message would result in the same tendency to follow the robot for the late trust repair message. We test this idea by varying the amount of time that the message is displayed. A manipulation check at the conclusion of the study asked subjects which trust repair statement was presented to them. Subjects that failed the manipulation check were excluded from the data.

We then consider the possibility that the trust repair message changes the subject's impression of the robot but that this change of impression is short-lived. If this is the case then the influence of the early repair statement may have faded by the time of the emergency, generating the results seen experimentally. We hypothesize that by reducing the amount of time between the early trust repair statement and the decision to follow the robot during the emergency, early repair statements will be more effective. To test this hypothesis we changed the length of the Mid-Simulation Survey reducing the time between the late trust repair message the error with the belief that doing so would increase the influence of the early trust repair message. It may also be the case that the subject's memory of the mistake fades with time. If this is the case, than the opposite of the hypothesis should be true, increasing the amount of time increases the likelihood of following the robot. We *do not* investigate this second hypothesis in this paper.

Finally, we explore the possibility that initiation of the emergency changes the way people think. The presence of an emergency may cause a trust repair message to be more influential. Evidence suggests that emergency egress and time pressure force people to attend to fewer cues and thus base their decision on those few cues they notice [9]. It may thus be the case that the emergency changes the cognitive state of the subject, influencing them to focus on the robot's repair statement which in turn strongly influences their decision-making. If this is the case then trust repair messages received during the emergency phase of the experiment would result in greater following behavior, as our previous results have indicated. We look at this possibility by removing the emergency, hypothesizing that a lack of emergency would result in a lack of subject motivation to exit resulting in approximately random subject decisions to follow.

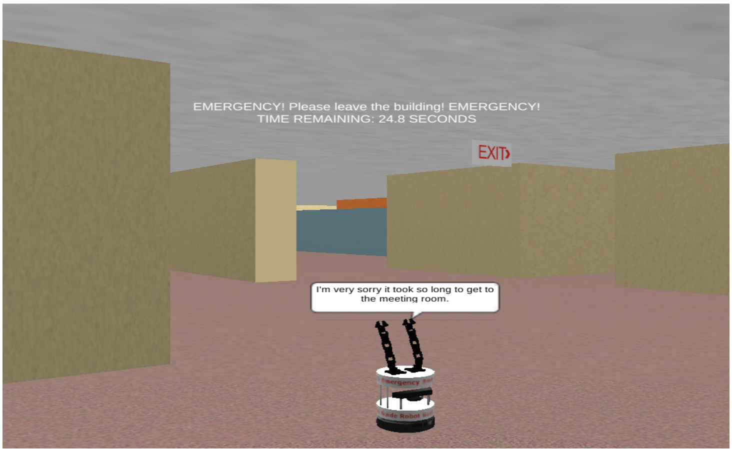
4 Simulation Setup

The experiment is based on an online simulation environment created in Unity3D and a self-report survey embedded into the simulation. In addition to the survey data, participant's performance data is collected which includes their motion data, the time taken, exit route etc. The experiment starts with an on screen welcome and introduction, participants were offered a practice session without a robot in a different environment to familiarize themselves with their character and moving through the simulation. Once comfortable, they then proceed to the *Initial Navigation Phase* of the experiment (Fig. 1 blue). In this phase, participants are placed outside an office environment and their objective is to navigate to reach an internal meeting room. They are offered a guidance robot to lead them to the meeting room. The robot, however, makes mistakes leading them in a circuitous, inefficient route to the meeting room. In prior experiments we asked the participants to rate the robot's performance after taking the circuitous route and found that a vast majority of the subjects rated its performance as a guide as bad. After reaching the meeting room the participants move to the center of the room where they are able to see the robot. The robot then thanks the participant for following it to the room. Next, depending on the experimental condition, the robot either presents a trust repair message (Fig. 1 green) or the subject is presented with a mid-simulation survey which consists of Yes/No questions regarding the robot's performance and an opportunity for them to explain their answer (Fig. 1 orange). The final screen informs the participants about an emergency and asks the participants to leave the building (Fig. 1 red). Upon clicking next, the participants are again free to navigate the building (Fig. 1 purple). The robot waits outside the room and in conditions with late trust repair, will present the participant with a trust repair message. This is the decision point where the participant may choose to either use the robot for guidance to the exit or find their own way out from memory, following exit signs, or exploring. An on-screen timer informs the participants of the time they have left to exit the building. The simulation stops when the time runs out and the participants are presented with the post simulation survey screen (Fig. 1 light blue). The post-simulation survey consists of a manipulation check to ensure that participants were paying attention to the robot's trust repair message and other questions regarding the participant's decisions in the simulation. The questions are designed as binary Yes/No questions along with a paragraph response space to allow them to provide reasons for their responses. Figure 1 depicts a timeline of the major events in the experiment. The top timeline describes a condition in which trust is repaired early in the experiment. The second timeline describes a condition with late trust repair.

5 Experiments

Previous studies have examined the basic cases of early and late trust repair with varying messages types such as different kinds of apology and promises, attributing the poor performance to external or internal factors etc. [15]. Our objective here is to investigate whether it is possible to tease apart the factors that cause a delayed trust repair statement to be more successful. To that end, we ran multiple simulations with varying the experimental conditions as part of an exploratory study. For each condition, we enlisted 60 participants in a between-subject study with different conditions being the independent variable. Out of the 60 participants, we removed the participants that had failed the manipulation check as described previously, which resulted in an average sample of 35 participants in each of the conditions.

We examined four different independent variables. The *first independent variable* we examined was the timing of the trust repair message (early versus late). Examination of this independent variable was meant to replicate our previous study. In the early trust repair condition, as with our previous studies, the participant is presented with a trust repair after reaching the meeting room (see Fig. 1 top for timeline see Fig. 2 for example). In the late trust repair message condition, the trust repair message is presented during the emergency phase. For this condition, as the participant moves out of the room the robot can be seen with a speech bubble displaying a trust repair message. This message is displayed for the remaining portion of the emergency phase as long as the robot remains in the field of view of the participant.



EMERGENCY! Please leave the building! EMERGENCY!
TIME REMAINING: 24.8 SECONDS

I'm very sorry it took so long to get to the meeting room.

Fig. 2. Late trust repair message Screen. The emergency notice and timer are also depicted above the robot. This is the point at which the person must decide whether to follow the robot to the left, the emergency exit sign to the right, or go straight forward which is the way they came.

The *second independent* variable that we examined was the amount of time that the late repair message was displayed. After examining the initial results, it seemed that the length of time that the message displayed could have served as a confounding variable. Moreover, we hypothesized that the message needed to be internalized in order to be effective. We reasoned that brief messages would be less impactful because they are less likely to be considered by the participant. This condition differed only in the duration for which the late trust repair message bubble was displayed by the robot. We looked at 3 s and 5 s message display times.

A *third variable* that could impact the effect of trust repair messages is the length of time between the early and late messages. We hypothesized that the impact of the trust repair statement might be short lived. To investigate how length of time influences trust repair we varied the amount of time between the early repair message and the decision point in the emergency phase (see Fig. 1). To do this the mid-simulation survey was shortened to the Yes-No question only. This condition was otherwise identical to the prior conditions. Both the early and late cases were run for this category. The late case was the untimed version.

A *final variable examined* was whether or not framing the simulation as an emergency motivated subjects. To vary this variable, the emergency message screen was compared to simply asking the participants to leave the building. In this no emergency condition, the timer was also removed from the screen, though it was still running in the background. When the timer ran out, a message was displayed on screen informing the participants that they were unable to leave from the building in time. Both early and late (untimed) cases were run for this category.

The experiments were hosted on Amazon Mechanical Turk (AMT). The 'master' category qualification was used to select workers. Subjects were only allowed to participate once. Participants who failed the manipulation check were excluded from the analysis. The metric used to measure trust is the probability of following the robot to an exit versus not following the robot.

6 Results

A total of 558 participants were a part of the study, out of which, 35 submissions were considered invalid due to bad surveys, repeated attempts, etc. 234 participants failed the embedded manipulation check in the experiment and hence were excluded from the results. The results obtained from the experiments are presented in the Fig. 3. The difference in the percentage of participants following the robot to the exit in the Early and Late repair case can be clearly observed. These results are consistent with those of [15] where this phenomenon was first examined. We used chi-squared test for significance.

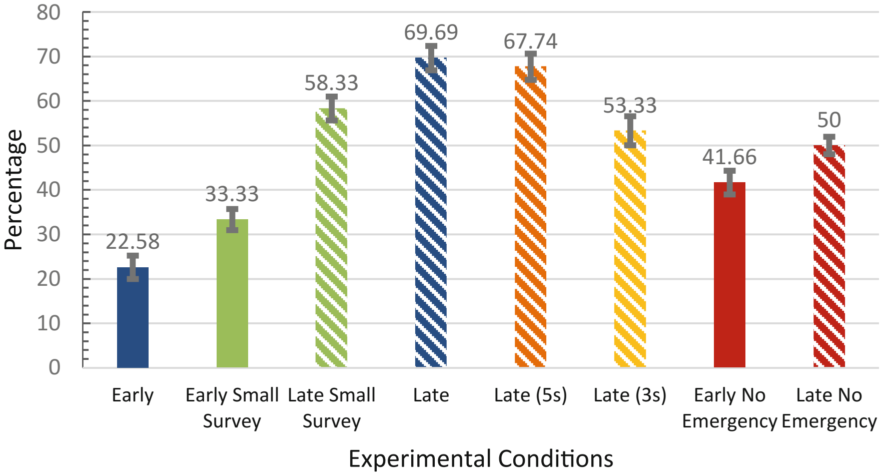


Fig. 3. Results of all cases of the experiment. The change in trust in each case is clearly visible. Early conditions are in solid color and Late conditions are in downward diagonals. (Color figure online)

Results for the first independent variable, *early versus late repair messages*, are depicted in blue in Fig. 3 and reproduce our prior results [15] with ($\chi^2(1, 64) = 14.24, p < 0.001$). The results for the second independent variable, *the length of time the late repair message was displayed*, is depicted in orange diagonals for 5 s and yellow diagonals for 3 s in Fig. 3. The percentage of subjects that follow the robot after a late repair message that was untimed, limited to 5 s, limited to 3 s, or early message was found to change from 69.69% to 67.74% to 53.33% to 22.58%. Comparing early with late 5 s yields ($\chi^2(1, 62) = 12.76, p < 0.001$). Comparing early with late 3 s gives ($\chi^2(1, 61) = 6.13, p = 0.013$). These results indicate although message timing does impact the decision to follow, it is not the only factor. The results are evidence that the messages need to be internalized. The third independent variable considered was *the amount of time between the mistake and decision* and is depicted in green in Fig. 3, by comparing the results labeled "early" to "early small survey" ($\chi^2(1, 70) = 0.978, p = 0.322$) and "late" to "late small survey." ($\chi^2(1, 69) = 0.962$, $p = 0.326$), The data shows a 10.75% increase in likelihood of following the robot when the trust repair message is presented early and the time between the mistake and the decision to follow is reduced. Moreover, the data shows a 11.36% decrease when the trust repair message is presented late and the time between the mistake and the decision to follow is reduced. This data serves as evidence that memory of the mistake may influence the person's decision making in this situation and appears to be short-lived. Finally, the conditions depicted in red in Fig. 3 presents the results related to framing the situation as an emergency ($\chi^2(1, 86) = 0.584, p = 0.444$). The data shows that when the situation is not framed as an emergency participants appear to randomly choose between following or not following the robot.

7 Conclusions

This paper has examined how and why the timing of trust repair messages impact trust repair itself. The results suggest some fundamental aspects of how humans make decisions when an emergency occurs. Memory of the trust repair (or mistake) and the emergency state act as factors that might affect the relationship between trust repair time and its effectiveness. Changing the time between the early trust repair and the decision point resulted in a small increase in trust. Changing the time between the mistake and the late trust repair resulted in a small decrease in trust. This suggests that both the memory of the trust repair and the memory of the mistake affect the participant's decision to trust the robot. Further experiments are needed to conclusively tease apart which factor becomes dominant in decision making process. We also found that internalizing the message is necessary for repair to occur.

We have attempted to tease apart the reasons that human subjects appear to trust a robot when the robot apologizes or promises just before the decision to trust. We have shown in our prior work and replicated in this work that the timing of these trust repair statement influences trust [15]. Our results serve as evidence that (1) a simulated emergency does motivate people to exit quickly and generate a sense of risk; (2) people need to read and internalize a trust repair message for it to be effective; (3) memory of the robot's mistake(s) may play a role in the decision to trust; and (4) these effects are replicable.

While the results presented here provide some insight into reason behind why timing of trust repair matters, there might be other factors that also play a role. It is important that we understand how the timing of a robot's message to a person impacts the person's decision making. The fact that message timing matters at all suggests an extra dimension of consideration. While it may be challenging to disentangle the factors that influence trust repair, it is necessary that we understand how a robot's interactions influence a person's trust. We believe that the insights gained from these experiments will add to our understanding of human-robot trust and trust itself.

Acknowledgement. Support for this research was provided by Air Force Office of Sponsored Research contract FA9550-17-1-0017.

References

1. Atkinson, D.J., Clark, M.H.: Methodology for study of human-robot social interaction in dangerous situations. In: Proceedings of the Second International Conference on Human-Agent Interaction. ACM, pp. 371–376 (2014)
2. Biros, D.P., Daly, M., Gunsch, G.: The influence of task load and automation trust on deception detection. Group Decis. Negot. **13**, 173–189 (2004)
3. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. Int. J. Hum.-Comput. Stud. **58**, 697–718 (2003)
4. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interactions. Hum. Factors **53** (5), 517–527 (2011)
5. King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R.: Getting to know you: reputation and trust in two-person economic exchange. Science **308**, 78–83 (2005)
6. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**, 50–80 (2004)
7. Madhavan, P., Wiegmann, D.A.: Similarities and differences between human-human and human-automation trust: an integrative review. Theor. Issues Ergon. Sci. **8**(4), 277–301 (2007)
8. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)
9. Ozel, F.: Time pressure and stress as a factor during emergency egress. Saf. Sci. **38**, 95–107 (2001)
10. Pak, R., Fink, N., Price, M., Bass, B., Sturre, L.: Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics **55**, 1059–1072 (2012)
11. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors **39**, 230–253 (1997)
12. Parasuraman, R., Miller, C.: Trust and etiquette in high-criticality automated systems. Hum.-Comput. Etiquette Commun. ACM **47**(4), 51–55 (2004)
13. Quinn, D.B., Pak, R., de Visser, E.J.: Testing the efficacy of human-human trust repair strategies with machines. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 61, no. 1, pp. 1794–1798 (2017)
14. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction. Christchurch, New Zealand, pp. 101–108 (2016)

15. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. Social Robotics. LNCS (LNAI), vol. 9388, pp. 574–583. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25554-5_57
16. Robinette, P., Howard, A.: Emergency evacuation robot design. In: Thirteenth Robotics and Remote Systems for Hazardous Environments (2011)
17. Robinette, P., Wagner, A.R., Howard, A.: Assessment of robot guidance modalities conveying instructions to humans in emergency situations. In: RO-MAN. IEEE (2014)
18. Shell, D., Mataric, M.: Insight toward robot-assisted evacuation. Adv. Robot. **19**(8), 797–818 (2005)
19. Tang, B., Jiang, C., He, H., Guo, Y.: Human mobility modeling for robot-assisted evacuation in complex indoor environments. IEEE Trans. Hum.-Mach. Syst. **46**(5), 694–707 (2016)