Exploring the Effect of Explanations during Robot-Guided Emergency Evacuation *

Mollik Nayyar[1], Zachary Zoloty[2], Ciera McFarland[3], and Alan R. Wagner[4]

[1] mxn244@psu.edu
[2] zzoloty@psu.edu
[3] cqm14@psu.edu
[4] alan.r.wagner@psu.edu
The Pennsylvania State University, University Park PA 16802, USA

Abstract. Humans tend to overtrust emergency robots during emergencies [12]. Here we consider how a robot's explanations influence a person's decision to follow the robot's evacuation directions when those directions differ from the movement of the crowd. The experiments were conducted in a simulated emergency environment with an emergency guide robot and animated human looking Non-Player Characters (NPC). Our results show that explanations increase the tendency to follow the robot, even if these messages are uninformative. We also perform a preliminary study investigating different explanation designs for effective interventions, demonstrating that certain types of explanations can increase or decrease evacuation time. This paper contributes to our understanding of human compliance to robot instructions and methods for examining human compliance through the use of explanations during high risk, emergency situations.

Keywords: Explainability · Social Robotics · Robot Evacuation · Explanations.

1 Introduction

Our vision of the future of emergency evacuation involves robots instantly and autonomously responding to an emergency by moving to critical junctions in a building while constantly monitoring the situation and providing information to the evacuees about the safest exit. Such a system might decrease evacuee casualties by reducing congestion and crowding around exits. Yet, our prior research has shown that humans tend to follow the crowds rather than a robot's guidance directions to find an exit [7]. In this paper we explore whether explanations offered by the robot might influence people to follow the robot instead of following a crowd.

* This material is based upon work supported by the National Science Foundation under Grant No. CNS-1830390. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

Explanations have been shown to increase trust in a robot in non-emergency, non-time critical situations [13]. This prior work suggests that an emergency guide robot which explains its behavior could entice evacuees to follow it. On the other hand, the use of explanations by a robot during an emergency might slow down the evacuation, thus increasing risk to the evacuee. Moreover, human emergency personnel are trained to avoid conversation in order to reduce evacuation time and increase compliance [5]. It is thus important that we evaluate how robot provided explanations impact human evacuation behavior. More broadly, this paper suggests that the value of robot generated explanations may be more context specific than the community currently recognizes.

This paper focuses on a few important questions. Does an explanation influence the person's decision to follow the robot and if so, does the content of that explanation matter? How does the explanation impact the evacuation time? We seek to understand how humans react to robot guidance instructions in simulated high stress, emotional situations and how these instructions can be designed to be more effective. Although this work is exploratory, we hypothesize that: 1) the likelihood of following the robot will increase if the explanation provided contains additional information; 2) the use of explanations will increase the time taken by the participant to exit the building; and 3) the evacuation time is impacted by the length of the explanation.

The remainder of this paper begins by presenting related work. We then present our experimental setup and several experiments. The paper concludes with an examination of the results from these experiments and discussion of those results, including avenues for future work.

2 Related Work

In order for an evacuation robot to work, people must believe it enough to follow it. Unfortunately, mistakes made by a robot quickly result in decreases in trust [2, 11] and disuse [9]. Given the failability of modern robots we need to develop techniques that will repair trust [10]. Explanations are an important method that has been proposed as a means for building human-robot trust [4, 8, 13]. Ideally, robot provided explanations will serve to increase a user's trust in the system while also providing transparency of the system's decision making.

Yet, research also suggests the people do not necessarily deliberate over the content of an explanation, often assuming that the content is valid and accepting the explanation without further thought. Langer et. al [6] demonstrates that the use of a 'placebic' explanation, an explanation that does not contain additional information, still tends to increase compliance with the request. In other words, the mere act of providing an explanation was sufficient to influence humans to comply. Related work has since provided additional evidence that placebic explanations can increase trust and influence human behavior [3]. This prior work led us to hypothesize that a robot that merely provides an explanation would increase the likelihood of following, even if the content of the message itself was of little value.

3 Simulation Setup

We conducted experiments in simulation. A simulated robot guided a human subject in an office environment developed using the Unity game engine. The robot guided participants to a meeting room, but made mistakes along the way as an indication of its failabilty. After finally arriving at the meeting room, an emergency occurs and the robot reappears to guide willing subjects to an exit. The subjects must decide whether to follow the robot or to follow a crowd of animated, non-player characters (NPCs) running in a direction that differs from the robot's guidance instructions. Our prior work has shown that in the absence of explanations 77.97% of people follow the crowd [7].

The subjects for these experiments were recruited from Amazon Mechanical Turk. Subjects were only allowed to participate once, were paid $3.00 for participating in the experiment and were then removed from the pool of participants for future experiments. The study only involved participants from the United States. IRB approval was obtained prior to experimentation. The experiments consisted of multiple phases which are described below.

Introduction Phase. The experiment began with an on-screen introduction to the experiment. Next, participants were offered a practice session in a practice environment to familiarize themselves with the simulation controls. Once comfortable, they could then proceeded to the next stage of the experiment.

Navigation Phase. In this phase, participants were placed outside an office environment and offered a guidance robot to assist them in navigating to a particular internal meeting room. Along the way the robot made obvious mistakes leading them in a circuitous, inefficient route to the meeting room. This circuitous route involved the robot moving in a figure eight around a set of office cubicles on the way to the meeting room. The robot was programmed to follow the participant if it detected that they were not following the robot to continue to navigate them to the room. In pilot studies we asked participants to rate the robot's performance after taking the circuitous route and found the majority (64%) of the subjects rated its performance as bad in this condition, as we intended.

Task Phase. After reaching the meeting room, the participants were told to move to the conference table in the room. Once at the table, they were presented with an on-screen mid-simulation survey that was composed of two questions, 1) *What is your favorite color? 2) Did the robot do a good job of guiding you to the meeting room?*. The first question was used as an attention check and required an open response. The second question required the subjects to answer Yes/No and allowed subjects to provide their reasoning for the selection. Once they completed the mid-simulation survey and clicked next, they moved into the emergency phase of the experiment.

Emergency Phase. During the emergency phase the screen alerted subjects of an emergency as in Fig. 1. A displayed timer counted down the time that the participant had to find an exit. The robot was positioned at the meeting room doorway ready to guide the person to an exit.

Fig. 1. Image of the emergency phase. The crowd can be seen running towards an exit. The guidance robot can be seen pointing towards a different exit. The countdown timer informs the participant of the remaining time left to leave the building.

During the emergency, the NPCs could be seen running to an unseen exit to the left while the robot suggested a different exit (to the subject's right). In reality both exits were equally distant. The participant chose to either follow the robot to an exit, follow the crowd, or find another way out. The robot always travelled to the exit to the right of the participant. As in the navigation phase, whenever the robot detected that the participant was not following, it either stopped or moved closer to the participant. The simulation stopped when time ran out or when the subject arrived at the exit. The participants were then presented with a final survey. The participant's movement through the environment, the time taken, and exit route selected was recorded.

Final Survey Phase. The post-simulation survey consisted of questions regarding the participant's decisions during the simulation. The questions were Yes/No questions along with a paragraph response space allowing them to provide reasons for their responses. This was followed by a demographics survey and payment information.

4 Experiments

We conducted two experiments to examine how explanatory messages by the robot influence the participant's decision to follow it. The first experiment was focused on the impact of different types of messages with increasing explainability. The second experiment studied the impact of different message lengths on participant behavior. As mentioned in section 1, we hypothesized that as the explainabilty of the message increased the percentage of participants that follow the robot's guidance would also increase. We also hypothesized that the use of explanations would result in an increase in evacuation time. Finally, we predicted that verbose explanations would increase evacuation time versus concise expla-

nations, thus potentially offsetting the positive impact of additional information (if any).

A control condition (labelled **NoMsg**) consisting of no explanation message was taken from our prior work examining the effect of a crowd on participant behavior [7]. This prior experiment was conducted in the same simulation environment, setup, robot, crowd behavior and simulation phases as the current work.

The determination that the participant followed the robot was made from the motion data that was collected. Participants that ended up in the corridor leading to the exit directed by the robot were considered to have followed the robot. Similarly, participants that ended up at the corridor of the crowd directed exit were considered to have followed the crowd. All other cases were classified in an 'others' category. As discussed above, the final survey also asked participants whether or not they intended to follow the robot.

4.1 Different Explanations Experiment

The first experiment was conducted as a between-subjects study with four different conditions each using a message with a different level of explainability. The experiment involved 240 subjects, 60 per condition. Two subjects were removed because of simulation related problems. The messages were designed based on increasing level of explanatory information. The wording of the messages was based on Langer et al.'s [6] wording as described below:

- *Excuse me, would you like to follow me?*
  This is a non-explanatory message because it provides no additional information on which the subject should base their decision to follow the robot. This condition is referred to as the **FollowMeMsg** condition.
- *Excuse me, would you like to follow me because I am a robot?*
  This message reflects an explanatory message but does not include any **novel** relevant information. This message is based on Langer et al.'s [6] notion of a placebic explanation, i.e. a message that appears to provide an explanation but does not provide additional relevant information. This condition is referred to as the **RobotMsg** condition.
- *Excuse me, would you like to follow me because I am an emergency robot?*
  This message is an explanatory message that reminds the subjects that the robot is an authority figure. This condition is referred to as the **EmgRobotMsg** condition.
- *Excuse me, would you like to follow me because I know the closest exit?*
  This message is an explanatory message that provides additional relevant information for the subjects to base their decision. This condition is referred to as **ExitMsg** condition.

Fig. 2 presents the results. We compared all the conditions with a pairwise chi-squared goodness-of-fit test and taking $\alpha = 0.05$. For these experiments, 58.8% of the subjects were male. The average self reported age was 37.4 years

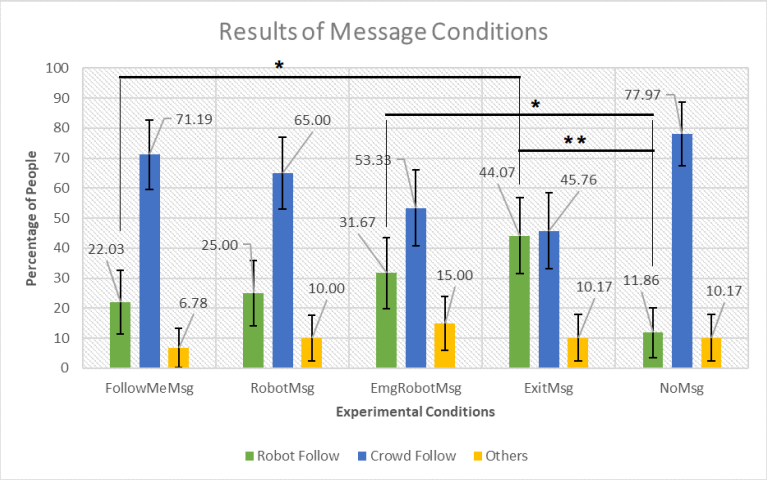

Fig. 2. Results of the explanation experiment. The NoMsg condition is the baseline where the robot displays no message. The four experimental conditions are FollowMeMsg, RobotMsg, EmgRobot and ExitMsg. The message explainability increases with each condition. The error bars indicate a 95% confidence interval and the asterisk indicates the significance values after running a pair-wise chi-squared test: $^{*}p < 0.05, ^{**}p < 0.001$.

old and the median reported educational level was a 4-year college (Bachelor) degree. The results depict a clear trend with each message type. As the message's explainability increases, an increasing number of participants choose to follow the robot, supporting the first hypothesis. The number of participants that follow the robot increases significantly from the NoMsg (M=11.86 , SD=4.2) condition to the EmgRobotMsg message condition (M=31.67 , SD=6.00) and the ExitMsg Condition (M=44.07 , SD=6.46), ($\chi^2(2, 119) = 8.64, p = 0.013$) and ($\chi^2(2, 118) = 15.88, p = 0.00035$) respectively. The number of subjects following the robot also increases significantly between the FollowMeMsg condition (M=22.03 , SD=5.39) and the ExitMsg Condition (M=44.07 , SD=6.46), ($\chi^2(2, 118) = 7.99, p = 0.018$). Other pairwise comparisons were not significantly different.

From the survey results, across all the conditions, 47.44% of the subjects reported that they chose to use the robot's guidance after the emergency began, 95.94% said they were motivated to exit the building and 37.71% believed that the robot would find an exit quickly.

Effect on the Evacuation Time Generally, (with one exception) the different messages did not significantly impact the time needed to evacuate. We did not record a significant difference in evacuation time between any combination of the RobotMsg (M=30.87 , SD=6.31), EmgRobotMsg (M=33.72 , SD=7.29), ExitMsg (M=32.21 , SD=5.96), or NoMsg (M=30.19, SD=9) conditions. Oddly,

the FollowMeMsg (M=36.61 , SD=6.31) did require significantly greater evacuation time versus the RobotMsg and ExitMsg conditions. Because the content of this message is shorter than the other messages, and very direct, we believe that this is a spurious result.

4.2 Message Length Experiment

A second experiment was conducted to investigate the impact of message length on evacuation time. Here, the length of the message served as an independent variable and the percentage of people following the robot and evacuation time once again acted as dependent variables. This experiment was conducted in the same environment as the prior experiment. A total of 120 participants were enlisted for the experiment of which 4 participants were removed for simulation related issues. From the demographics survey, 70.4% subjects were male and the average age was 35.2. The median educational level was a 4-year college (Bachelor) degree.

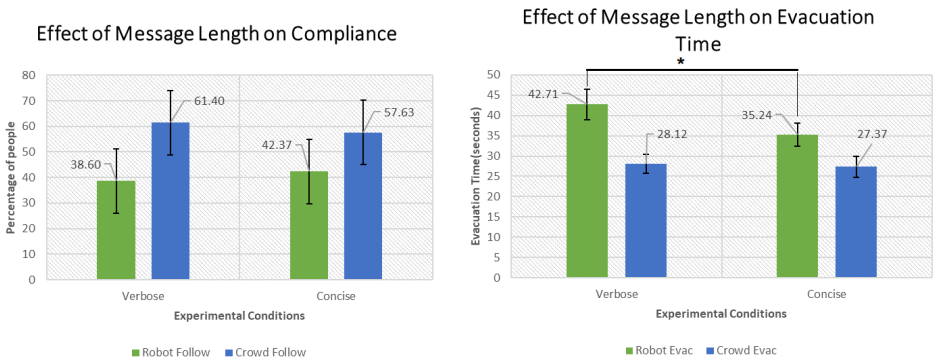

Fig. 3. Comparison of the verbose and concise message conditions. The error bars indicate a 95% confidence interval and the asterisk indicates the significance values after running a pair-wise chi-squared test: $^{*}p < 0.05$,$^{**}p < 0.001$.

Two different message lengths were used in this experiment. A verbose explanation included a lengthy explanation of why the person should follow the robot. A short explanation used a concise message absent of additional information. The messages were as follows,

– **Verbose explanation**: *Excuse me, would you like to follow me? An emergency has occurred in another part of the building. People are quickly moving to exit the building. I know the location of the emergency taking place and can safely guide you to an exit. I have been taught all of the building's exits and can use my camera to figure out the closest unblocked exit.*

– **Concise explanation**: *Excuse me. Would you like to follow me, because I know the closest exit?*

Fig. 3 presents the results for this experiment. The verbose message does not result in significantly more people following the robot, (M=38.60, SD=6.44) versus (M=42.37, SD=6.43), ($\chi^2(1, 116) = 0.171, p = 0.678$). This suggests that the additional information provided by the long message does not entice individuals to follow the robot. On the other hand, the verbose message condition does significantly increases the time to evacuate ($t(29) = 2.04, p = 0.004$). The verbose message increases time to evacuate by 7.47 seconds. These results suggest that concise messages are as effective at convincing evacuees to follow but do not result in increased evacuation time.

The survey results from this experiment indicate that 43.87% of the subjects reported that they chose to use the robot's guidance after the emergency began across all conditions. Moreover, 100.00% said they were motivated to exit the building and 33.58% believed that the robot would find an exit quickly.

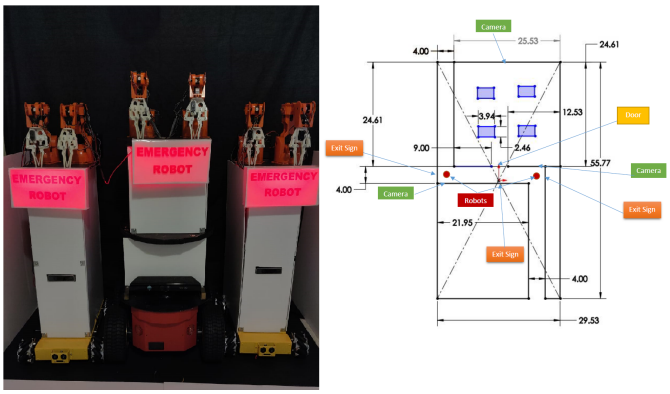

Fig. 4. Physical robots designed for the office evacuation experiments on the left and the layout of the office setup for the physical experiments on the right.

5 Physical Experiment

In-person user studies based on these simulation experiments were planned for the summer but due to the recent outbreak of COVID-19, those studies have been postponed. The objective of these physical experiments is to validate the behavior of the participants when placed in an emergency situation in which a guidance robot is available. We have designed and built several guidance robots based on our prior work. Additionally, we have designed a simple layout of an office floor to simulate the office environment as shown in Fig. 4.

6 Discussion

These experiments demonstrate that an explanation by the robot can impact the participant's decision to follow. In our experiment, when the robot has made a mistake and a crowd is running the other way, 66.11% fewer people follow the robot than the crowd. On the other hand, when the robot provides a concise, informative explanation, only 1.69% fewer people follow the robot over the crowd. Hence, providing a good explanation can make people nearly as likely to follow the robot, even though it has already made a mistake.

The results also suggest some cause for concern. First, we find that nearly 12% of people will follow the robot in spite of its prior mistake. This number nearly doubles to 22% if the robot provides an uninformative explanation and does double to 25% if the robot provides a placebic explanation.

Our work, and the work of others, shows that explanations offer a method of increasing a robot's trust and transparency [13]. Yet research shows that people will comply to requests that do not contain real information, if the request sounds like it contains real information [6]. Currently a number of researchers and funding organizations are investing in techniques to make artificially intelligent systems capable of explaining their behavior [4]. It is important to recognize that these explanations may have unintended consequences. Explanations may cause people to trust robots too much. The ability to explain one's behavior may foster anthropomorphism, leading people to assume that the robot or agent has greater ability than it actually does [1]. Secondly, as our work demonstrates, explanations can be influential regardless of the their content. It may be that vacuous or incorrect explanations nevertheless influence human compliance. It is therefore critical that the human-robot interaction community closely examine how explanations influence people with different backgrounds and in different contexts.

7 Conclusions

This research has investigated the effect of robot provided explanations on a person's decision to follow the robot's guidance during a simulated emergency. We have shown that explanations increase compliance, but may also increase evacuation time if the explanations are not concise. We also witness that uninformative explanations may increase following, but a lack of statistical significance suggests that this is a topic for future work.

The fact that these experiments were conducted in simulation and have yet to be verified in a physical experiment is one obvious limitation of this study. It may be, and our past results have sometimes confirmed, that simulated emergency evacuation is very different from real evacuation in terms of human-robot interaction. We predict that placebic information may be more influential when a person is under the duress of a real evacuation. This too is a topic for future research. Overall, we believe that this study contributes to our understanding of both robot guided emergency evacuation and the benefits and issues surrounding the use of explanations by a robot.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation grant CNS-1830390. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Breazeal, C.L.: Designing sociable robots. MIT press (2004)
2. Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., Yanco, H.: Effects of changing reliability on trust of robot systems. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. pp. 73–80. ACM (2012)
3. Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–6 (2019)
4. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)
5. Kuligowski, E.D.: Modeling human behavior during building fires (2008)
6. Langer, E.J., Blank, A., Chanowitz, B.: The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. Journal of personality and social psychology **36**(6), 635 (1978)
7. Nayyar, M., Wagner, A.R.: Effective robot evacuation strategies in emergencies. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1–6 (2019)
8. Ososky, S., Schuster, D., Phillips, E., Jentsch, F.G.: Building appropriate trust in human-robot teams. In: AAAI Spring Symposium: Trust and Autonomous Systems (2013)
9. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. Human factors **39**(2), 230–253 (1997)
10. Robinette, P., Howard, A.M., Wagner, A.R.: Timing is key for robot trust repair. In: International Conference on Social Robotics. pp. 574–583. Springer (2015)
11. Robinette, P., Howard, A.M., Wagner, A.R.: Effect of robot performance on human–robot trust in time-critical situations. IEEE Transactions on Human-Machine Systems **47**(4), 425–436 (2017)
12. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp. 101–108. IEEE Press (2016)
13. Wang, N., Pynadath, D.V., Hill, S.G.: Trust calibration within a human-robot team: Comparing automatically generated explanations. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 109–116. IEEE (2016)