Better Faulty than Sorry: Investigating Social Recovery Strategies to Minimize the Impact of Failure in Human-Robot Interaction

Sara Engelhardt, Emmeli Hansson, Iolanda Leite

Department of Robotics, Perception and Learning
KTH Royal Institute of Technology

Abstract. Failure happens in most social interactions, possibly even more so in interactions between a robot and a human. This paper investigates different failure recovery strategies that robots can employ to minimize the negative effect on people's perception of the robot. A between-subject Wizard-of-Oz experiment with 33 participants was conducted in a scenario where a robot and a human play a collaborative game. The interaction was mainly speech-based and controlled failures were introduced at specific moments. Three types of recovery strategies were investigated, one in each experimental condition: ignore (the robot ignores that a failure has occurred and moves on with the task), apology (the robot apologizes for failing and moves on) and problem-solving (the robot tries to solve the problem with the help of the human). Our results show that the apology-based strategy scored the lowest on measures such as likeability and perceived intelligence, and that the ignore strategy lead to better perceptions of perceived intelligence and animacy than the employed recovery strategies.

1 Introduction

Social interactions are not always successful, but humans have developed social norms to deal with such cases. Sunstein defined social norms as "social attitudes of approval and disapproval, specifying what ought to be done and what ought not to be done"[6]. Considering that even interactions between humans fail at times, it is not surprising that Human-Robot interaction (HRI) might inevitably fail too, especially because of some robot malfunction. These failures can be critical because they might require costly human intervention and, more importantly, they can cause users to lose trust and interest in the robot.

Giuliani et al. found two types of failures in HRI: technical failures and social norm violations [4]. While technical failures are often a result of execution failures (i.e., an appropriate action that was carried out incorrectly), social norm violations are defined as "a deviation from the social script or the usage of the wrong social signals" [4]. Social norm violations often occur due to planning failures or actions that are executed correctly but are inappropriate to the situation. An example of a planning failure would be the robot asking the user the same question several times even though an appropriate answer has been given. Inappropriate social signals can occur, for example, when the robot is not looking at the person it is talking to [4].

This paper will investigate the impact of social norm failures committed by a robot and how that impacts people's perception of the robot. We limit the scope of our study to verbal failures during human-robot conversation because we anticipate that speech recognition failures will be one of the main causes of disruption of the natural course of the interaction once social robots are deployed in real world environments. In a between-subject pilot study, we investigated people's perceptions of a robot that employed one of out three types of failure recovery strategies (ignore, apology or problem-solving).

2 Related Work

2.1 Human Perception of Robot Failure

Earlier work has studied how a robot is perceived by the user in failure situations. In a study by Lee et al. [5], where the interaction consisted of a human asking the robot to get a drink and receiving either the right or wrong one, it was found that robot failure decreased all ratings of the robot compared to the successful interaction, except how much they liked the robot. On the other hand, Bajones, Weiss, and Vincze [1] found a tendency for only a small negative impact on perceived intelligence, likability and robot contribution when a robot malfunctions, in their study about how to mitigate robot failure with the help of the user. They attributed that to the robot's recovery strategies that made it able to fulfill its tasks in the end. The importance of task success is also found in a study by Foster et al. [3], where a robot is used as a bartender. They found that dialogue efficiency and task success had the biggest impact on the subjective measures, as well as perceived intelligence and likeability, which showed a generally positive outcome [3]. However, Bajones, Weiss, and Vincze [1] noted that repeated demands for help became an annoyance. Torrey, Fussell, and Kiesler [7] studied how a robot is perceived when offering advice. Using hedges and discourse markers improved how considerate, controlling and likable the participant perceived the robot. Their results also indicate that robots using politeness might have an even bigger positive effect on the interaction than humans doing the same. Lee et al. [5] also concluded that the importance of politeness is apparent (politeness ratings of the robot increased with every recovery strategy used).

2.2 Failure Recovery Strategies in HRI

When a robot fails, just like when a human fails, there are different ways to recover from that failure. Earlier studies have tested different recovery strategies, for example, the robot stating what the problem is and how the participants could help fix it when it malfunctions [1]. This recovery strategy has a *problem-solving* approach, where the robot asks the users for help when necessary. The aforementioned study by Lee et al. [5] compared four different recovery strategies: *forewarning*, where the robot warns that it might fail at the start of the interaction; *apology*, where the robot apologizes for the failure; *compensation*, where the robot offers some kind of compensation for the failure; and *option*, where the robot suggests different ways to try to solve the failure. In their control condition, the robot's only response to the failure was to say "OK". Overall, the apology strategy scored best for them. However, people with low relational or high utilitarian orientation liked compensation best, and actually preferred no recovery strategy over both apology and options, suggesting that recovery strategies need to be tailored to a persons orientation to services, and different scenarios might benefit from different strategies [5]. It is important to note that the findings by Lee et al. were obtained in the context of service robots. It remains unknown whether similar results will hold for other types of human-robot interaction, like the collaborative scenario described in this paper.

3 Method

The research question we seek to answer with this pilot study is the following: which is the optimal strategy that robots can use to minimize the negative impact of failure in social collaborative interactions with a human?

3.1 Scenario

We investigated social interaction failure in a collaborative task between a person and a humanoid robot Nao[1] as displayed in Figure 1. Twelve cards were placed on a table, facing down, in spots labeled from A to L. The goal of the task is to find all the four queen cards, the robot already having the first one. Because the robot is not capable of turning the cards itself, it asks for the human's help to turn certain cards (one at a time, using the labels) and say which card is where, in order to find the hidden queens.

[1] https://www.ald.softbankrobotics.com/en/cool-robots/nao

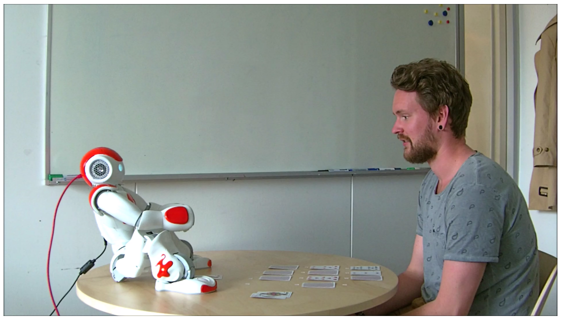

Fig. 1. The experiment setup with the robot, Nao, and a participant.

The failure consisted of the robot interpreting human speech incorrectly when asked about which card is in a specific location. To make it clear that the robot fails, the robot always repeats the card it just heard and asks the participant to confirm or deny. When the participant answers "no" to this question, different recovery strategies will be used depending on the experimental conditions. Each recovery strategy will have its own protocol of how to handle and recover from the failure (more details in Section 3.2).

Since the main focus of the study was to investigate failures, making sure the failures were controlled and the same for every experiment was important. Therefore, a Wizard-of-Oz-technique was used for the verbal aspects of the interaction because the speech recognition for the robot is not reliable enough yet. The non-verbal aspects, such as face tracking and gestures of the robot were fully autonomous and not synchronized to the game.

3.2 Conditions

The three strategies for mitigating the negative impact of failure in social contexts that we want to study are based on the strategies used previously in the work of Lee et al. [5]:

- **Ignore**: the robot ignores that it has failed and simply keeps going by saying "OK". This can be considered our control condition.
- **Apology**: the robot apologizes (for example, by saying "I'm sorry, sometimes I don't interpret speech correctly") for its failure and then moves on.
- **Problem-solving**: the robot tries to solve its failure with the help of the human by asking him/her to repeat the card. After hearing the card one more time, the robot acknowledges that it understood which card the participant is referring to.

3.3 Procedure

Participants were guided to a room by an experimenter and instructed to sit at a table in front of the robot. The robot then provided the instructions for the game (find the remaining queens). When 9 of the 12 cards were flipped the game would finish. The outcome of the game would always be the same, for example, all queens were found. Also, to make sure the outcome of the game was not affected, all failures occurred in cards other than queens. Each participant experienced three failures, three queens found and heard correctly and three other cards that were heard correctly. For all conditions, failures occurred on cards number 3, 6 and 7, while the other 6 plays were normal interactions (i.e., no failure).

The average length of the interaction with the robot across all conditions was about 3.0 minutes, with the average for the ignore condition slightly shorter (M=2:49, SD=15sec) and for problem-solving slightly longer (M=3:13, SD=32sec). The apology condition had the same average as the overall time (M=3:01, SD=17sec). At the end of the interaction, the robot thanked the participant for playing, and the participant was asked by the experimenter to fill in a questionnaire.

3.4 Measures

At the end of the interaction, participants filled in a survey consisting of two parts: some general questions about the participant and perceptual measures commonly used in HRI experiments. The general questions included participant's age, gender, occupation, as well as experience with programming and experience with robots on a five-point Likert scale for Godspeed and nine-point for RoSAS.

The perceptual measures were taken from the Godspeed Questionnaire [8] and the The Robotics Social Attributes Scale (RoSAS) [2]. The Godspeed Questionnaire Series is an established way to measure peoples perception of robots. There are five measures in total: Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety. We chose to only use **Animacy**, **Likeability** and **Perceived Intelligence** since we believe these are the most relevant measures to our study. The Robotics Social Attributes Scale (RoSAS) builds upon the Godspeed Questionnaire Series, but seeks to improve the cohesiveness of the measurements [2]. We used **Competence** and **Discomfort**, but chose to exclude the warmth metric because it seemed less relevant to this study. All the different items of each measure were randomized in the survey.

3.5 Participants

We recruited 33 adults participants for our study (11 participants per condition). All participants were computer science undergraduate students from the same university in Sweden and their median age was 22 years old. There is a majority of male students in computer science programs, which was reflected in the gender distribution of our participants. Participants were randomly distributed between conditions. In the ignore condition we had 64% males (7) and 36% females (4); in the apology condition we had 82% males (9) and 18% other (2), and in the problem-solving condition, we had 73% males (8) and 27% females (3). Participants reported that their previous experience with robots was low, with an average of 1.8 on a scale from 1 to 5.

4 Results

4.1 Likeability

Generally the participants in all three conditions rated the robot high on likeability. Participants in the apology condition rated the robot slightly lower ($M = 4.3, SD = 0.4$) than in the ignore ($M = 4.5, SD = 0.7$) and problem-solving condition ($M = 4.5, SD = 0.2$).

4.2 Perceived Intelligence

In general, the robot was perceived as fairly competent and intelligent, with average scores mostly between 3 and 4. The ignore condition scored highest ($M = 3.7, SD = 0.6$), followed by problem-solving ($M = 3.4, SD = 0.4$), and apology ($M = 3.1, SD = 0.6$).

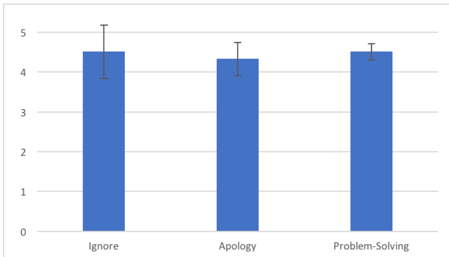

Fig. 2. Results for Likeability.

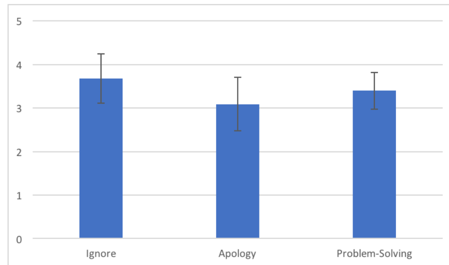

Fig. 3. Results for Perceived Intelligence.

4.3 Animacy

Overall, the robot was seen as fairly animated, with scores slightly above the middle of the scale. The ignore condition generated a slightly higher rating ($M = 3.4, SD = 0.7$) than apology ($M = 3.2, SD = 0.6$) and problem-solving ($M = 3.2, SD = 0.4$).

4.4 Competence

Generally, the apology condition ($M = 5.3, SD = 1.3$) rated the robot lower on perceived competence, with the ignore condition ($M = 6.2, SD = 0.9$) scoring slightly higher than problem-solving ($M = 6.0, SD = 1.2$).

4.5 Discomfort

The robot scored low on discomfort, with the averages for all three conditions below 3 on the 9-point scale. Ignore ($M = 3.0, SD = 1.5$) scored slightly higher than apology ($M = 2.7, SD = 0.9$) and problem solving ($M = 2.6, SD = 0.9$).

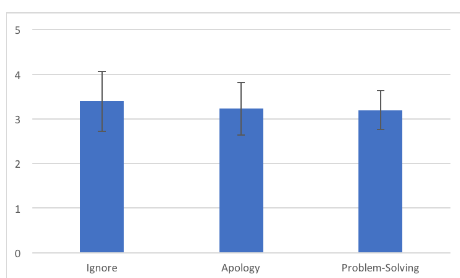

Fig. 4. Results for Animacy.

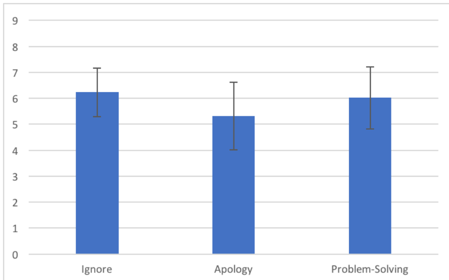

Fig. 5. Results for Competence.

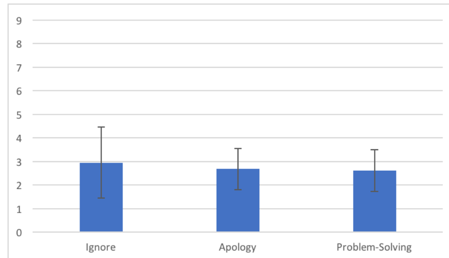

Fig. 6. Results for Discomfort.

5 Discussion and Conclusion

The low scores on discomfort and high scores on robot likeability can be explained by several factors such as the robot's appearance, the non-critical nature of the task and also the participants' background (computer science students). An interest in technology, experience with programming and some experience with robots could all reduce the feeling of discomfort in the presence of a robot, as well as increase the interest for, and liking of, robots.

An important aspect of all the participants in our study is that their education focuses heavily on problem-solving, which might attract students with a problem-solving mindset. This could be the reason why the participants in our study preferred the problem-solving strategy over the apology strategy. As Lee et al. [5] found, participant individual traits can affect which recovery strategy is preferred.

The ignore condition was designed to be less responsive and not intended to recover from failure. In the study by Lee et al. [5], a similar behavior resulted in a lower score for perceived competence compared to problem-solving and apologizing strategies. However, in our study, the apology condition scored lowest on intelligence and competence scores, while the ignore condition scored highest on responsiveness. This suggests that the ignore condition was not perceived as expected. We believe this might have happened due to an unexpected pattern in the participants behavior that emerged during the experiments; many participants instinctively repeated the card that the robot misheard. The ignore condition could be seen as acknowledging this when it says "OK", while the other conditions clearly ignored it. Not only does this make the robot seem more responsive in the ignore condition, but it can be perceived as the robot having corrected the mistake, similar to what happened in the problem-solving condition. That would leave apology the only one that does not correct the mistake, which could affect its ratings negatively.

In conclusion, based on previous research, we expected that the apology strategy would do best, followed by problem-solving and then ignore. However, our results show that the apology condition resulted in less positive perceptions of the robot in most of our measures of interest, and that the ignore condition did as well as, or better than, problem-solving in many cases.

5.1 Limitations and Future Work

Our participant pool was a very homogeneous group: all participants were between 20 and 27 years old and studied computer science at the same university. The rather small sample size entails that wizard or participant errors might have a larger impact on the results, but the homogeneousness of the group is an advantage when drawing conclusions about this particular age group.

There is still a lot of research to be done in this area. While the results of this pilot study were quite informative, a large sample size would enable us to apply statistical analysis to determine whether the trends we found are statistically significant. In the future, we should also consider including a condition where the robot does not fail at all, to see how much the failure itself influences the participants perception of the robot.

References

1. Bajones, M., Weiss, A., Vincze, M.: Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user (2016)
2. Carpinella, C., Wyman, A., Perez, M., Stroessner, S.: The robotic social attributes scale (rosas): Development and validation. In: Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction. pp. 254–262. HRI '17, ACM (March 2017)
3. Foster, M.E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., Petrick, R.: Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In: Proceedings of the 14th ACM international conference on multimodal interaction. pp. 3–10. ICMI '12, ACM (October 2012)
4. Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., Tscheligi, M.: Systematic analysis of video data from different human-robot interaction studies: a categorization of social signals during error situations. Frontiers in psychology 6 (2015)
5. Lee, M.K., Kielser, S., Forlizzi, J., Srinivasa, S., Rybski, P.: Gracefully mitigating breakdowns in robotic services. In: Proceedings of the 5th ACM/IEEE international conference on human-robot interaction. pp. 203–210. HRI '10, IEEE Press (2010)
6. Sunstein, C.R.: Social norms and social roles. Columbia Law Review 96(4), 903–968 (1996)
7. Torrey, C., Fussell, S., Kiesler, S.: How a robot should give advice. In: Proceedings of the 8th ACM/IEEE international conference on human-robot interaction. pp. 275–282. HRI '13, IEEE Press (2013)
8. Weiss, A., Bartneck, C.: Meta analysis of the usage of the godspeed questionnaire series. pp. 381–388. IEEE (August 2015)