Exploring the Impact of Fault Justification in Human-Robot Trust

Socially Interactive Agents Track

Filipa Correia
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
filipacorreia@tecnico.ulisboa.pt

Carla Guerra
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
carla.guerra@gaips.inesc-id.pt

Samuel Mascarenhas
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
samuel.mascarenhas@gaips.inesc-id.pt

Francisco S. Melo
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
fmelo@inesc-id.pt

Ana Paiva
INESC-ID & Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
ana.paiva@inesc-id.pt

ABSTRACT

With the growing interest on human-robot collaboration, the development of robotic partners that we can trust has to consider the impact of error situations. In particular, human-robot trust has been pointed as mainly affected by the performance of the robot and as such, we believe that in a collaborative setting, trust towards a robotic partner may be compromised after a faulty behaviour. This paper contributes to a user study exploring how a technical failure of an autonomous social robot affects trust during a collaborative scenario, where participants play the Tangram game in turns with the robot. More precisely, in a 2x2 (plus control) experiment we investigated 2 different recovery strategies, justify the failure or ignore the failure, after 2 different consequences of the failure, compromising or not the collaborative task. Overall, the results indicate that a faulty robot is perceived significantly less trustworthy. However, the recovery strategy of justifying the failure was able to mitigate the negative impact of the failure when the consequence was less severe. We also found an interaction effect between the two factors considered. These findings raise new implications for the development of reliable and trustworthy robots in human-robot collaboration.

KEYWORDS

Social Human-Robot Interaction; Technical Failure; Faulty Robots; Error Recovery; Recovery Strategy; Trust; Cooperation

ACM Reference Format:
Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 7 pages.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

1 INTRODUCTION

Robots, like any other machines, are susceptible to fail or present some degree of error. We are all familiar with a robot that suddenly halts, starts repeating itself, says something out of context, and many other situations. Depending on the nature of the task and the purpose of the robot, the impact of failures can range from amusing to highly dangerous. However, even in low-risk situations, such as a conversational, entertainment or companion robot, failures may have a significant adverse effect on trust, user engagement and even willingness to interact with the robot in the future. From a performance standpoint, if robots are able to understand and recover from their failures automatically, they will be more efficient and reliable. But as robots become more social and interact with humans in various forms, the expectations on how robots handle such failures may go beyond their capacity for autonomous recovery. For instance, in collaborative tasks where robots are interacting with humans, the robot's behaviour should also address the social implications of their failures upon others. If we expect others to justify and explain their failures to us, it is likely that we will expect social robots to do so as well. This work sets out to understand how a robot can recover from a failure in order to mitigate its possible negative social effects. In particular, if a robot justifies the failure, will it mitigate the effects of it? By addressing these questions, this work contributes to the design of social agents that can autonomously overcome error situations in a more appropriate manner.

Literature related with faulty or erroneous behaviours in social agents and robots is still very recent. As a result, the range of effects that is caused by such behaviours is not yet fully understood or agreed upon. For instance, Salem et al. found that incongruent "speech-gestures" lead the robot to be perceived as more anthropomorphic, human-like, and likeable when compared to congruent "speech-gestures" [14]. Conversely, Mirnig et al. found no significant difference between anthropomorphism or perceived intelligence between the faulty and the flawless robots, although the faulty robot was rated with higher levels of likeability [8]. In addition, there are also findings reporting negative effects on the perception

of a robot after faulty behaviours, as shown by Ragni et al. [10] or Salem et al. [15]. Nonetheless, the controversy of previous findings suggests that the perception of a robot after an error situation may be influenced by many factors, such as the type of task, the type of error, or the severity of the error. Naturally, it extends the scope of unexplored issues regarding the topic of faulty behaviours in social robots.

In addition, some researchers began to explore mechanisms to cope with faulty behaviours and, possibly, mitigate their negative effects, known as recovery strategies. The previously analysed recovery strategies [2, 7] were developed for robots in service tasks and were tested through online surveys where participants did not directly interact with the faulty robots. Moreover, if the perceptions of faulty robots seem to be influenced by several factors, the mitigation strategies must also be explored for different levels of the same factors.

Yet, in spite of these different findings, it is still unclear if justification of a fault by the robotic agent will have a positive impact in the people interacting with it. That is, if the robot makes the problem/fault transparent to the user and justifies it, will it mitigate the perception of severity of that fault? And the trust?

In this paper, we contribute to this emerging field by exploring the impact of fault justification as a recovery strategy, which, to our knowledge, has not yet been explored. Our contribution consists of a user study that was conducted with an autonomous social robot collaborating with participants in a shared task, a puzzle game. At a certain point, the robot has a technical failure during the task and, depending on the experimental condition, adopts a different social recovery strategy.

The motivation behind analysing trust is the fact that it is one of the most critical and essential elements for an effective collaboration between humans and agents [6]. Moreover, according to Hancock et al., trust is strongly influenced by the agent's performance and other attributes; such as transparency.

Overall, the obtained results indicate that the recovery strategy of justifying the failure was able to mitigate the negative impact of the failure, but only when the consequence of the failure was less severe (when the failure did not compromise the task). That is, in scenarios where the failure is not too severe, a strategy of justifying a failure to the users can mitigate the overall trust in the robot. The implications of these findings are particularly relevant to the current growing interest in collaboration between humans and agents, or in tasks where agents act as peers or constitute a team with humans.

This paper is organised as follows: first we will discuss the work being done in the Human-Robot Interaction (HRI) community concerning failures and their impact on humans; then, a definition of trust in HRI is given; afterwards, our scenario is presented in detail, followed by a description of the user study that was conducted using the previous scenario; the obtained results are then presented and discussed in detail; finally, we present our conclusions and discuss the implications of this work for the community.

2 RELATED WORK

The study of error situations in HRI is still new. Currently, there are three broad questions that are getting researchers' attention in this area: (1) How can a robot automatically perceive error situations? (2) How do error situations influence the interaction and human perception? (3) Which strategies can be adopted to mitigate the effects of a failure?

Giuliani et al. [3] have postulated that robots must be able to recognise social signals during error situations. Through a video analysis in different user studies, the authors have evaluated the social signals humans perform related to error situations during human-robot interaction. Their analysis shows that participants are prone to, for instance, using head movements and laugh when the error situation occurs but are not prone to using hand gestures. Another relevant consideration from their video analysis is that they could classify error situations in two distinct ways, namely, social norm violations and technical failures. The former type corresponds to error situations that provide the wrong social signals or produce a discrepancy in the social script. Differently, the latter type refers to failed attempts to perform an action.

Salem et al. have analysed the effects of robot gestures on the perception of the robot [14]. In one of their conditions, the erroneous behaviour of the robot is associated to incongruent multimodal behaviours (speech and gestures). Their findings revealed that the anthropomorphic perceptions and the mental models of the robots can indeed be influenced by the communicative non-verbal behaviours. Interestingly, the incongruent multimodal behaviour was rated with greater humanlikeness and likeability when compared to the congruent multimodal behaviour.

In another user study, Salem et al. have manipulated a robot to display either correct or faulty behaviours in the beginning of the interaction, and then ask unusual requests, e.g. dispose letter, pour orange juice, disclose information [15]. Although people perceived the faulty robot as less humanlike, reliable and trustworthy, the manipulation had no impact on their willingness to comply with the unusual requests. Similar findings by Robinette et al. reported a tendency to follow and overtrust a robot in an emergency evacuation scenario, regardless of the robot having previously displayed faulty behaviours or not [13].

During a competitive scenario by Ragni et al., a robot was manipulated to either produce some occasional mistakes and display limited memory, or to always perform correctly [10]. Although the erroneous robot was perceived as less intelligent and reliable, participants perceived the interaction as easier and more positive. An interesting result was that people interacting with the erroneous robot presented a lower performance of executing the task. The authors attributed this result to a calibration of performance set by the perceived performance of the robot.

However, perceptions of faulty behaviour are not consistent and, for instance, Mirnig et al. reported no significant differences in anthropomorphism nor in the perceived intelligence between a social robot performing correctly and faulty [8]. The authors attributed this result to the fact that the error was non-task related. Nevertheless, participants liked the faulty robot significantly more than the flawless robot, which points toward the idea that a faulty social robot can actually be perceived as more natural.

One last example where the interaction and perception of a faulty robot was investigated in a slightly different scope is by Sarkar et al., who conducted an experiment with a collaborative manufacturing task where the robot performed faulty or not [16]. Their results

show that the faulty condition compared to the non-faulty one did not have a significant effect on the social perceptions of the robot nor in their performance in the task. The authors attribute this result to the nature of the task, which was difficult and demanding.

Regarding the possible strategies that robots can employ after an error situation, there are two online user studies reporting that recovery strategies can indeed mitigate the negative impacts of robotic failures. Lee et al. showed that the apology strategy was most effective to mitigate perceptions of competence, closeness and likeability of a service robot [7]. However, the authors also showed that people's orientation to services may lead to different effects of the observed recovery strategies.

In a similar online survey, Brooks et al. explored people's reactions to failures in autonomous robots [2], namely a vacuum cleaner and a self-driving taxi, by manipulating four variables: context risk, failure severity, task support and human support. Participants' perceptions of an erroneous robot became less negative when it deployed a mitigation strategy, either by prompting task support, human support or both. However, the authors reported an interesting but non-significant tendency showing a preference for both task and human support in high severity situations, and a preference for only task support in low severity situations. Generally, Brooks et al. contributed to previous results of Lee et al. [7] with the notion that the amount the strategy influenced people's reactions depend on the type of task, the severity of the failure and the risk of the failure.

Overall, the studies previously mentioned reveal that there are several factors influencing the way people perceive a faulty robot, as severity of the erroneous behaviour [14] or the type of the task and context [10, 13, 16]. Moreover, there seems to be a degree to which faults are accepted, even considered as more positive and more likeable [8, 10] and can even calibrate human performance to regulate the trade-off between performance and satisfaction during the interaction [4, 10]. Nevertheless, even when failures negatively influence the perception of the robot or the interaction, there are also findings reporting the effectiveness of recovery strategies for mitigating the negative impact of failures [2, 7].

One of the disadvantages of conducting an online survey for the perception of a faulty robot is that participants act only as observers. As such, they are not directly affected by the robot's failures. Our work avoids this issue as the participants in our study rate their perceptions of the faulty robot after interacting with it as a peer and being directly affected by it's faulty behaviour. Another distinguishing aspect is that we look at the mitigation effects of an unexplored recovery strategy, namely, the robot justifying the fault. Finally, our robot acts autonomously and simulates an autonomous recovery of a technical failure, which differs from failing to accomplish a goal [7] or presenting erroneous behaviours during the interaction [15].

3 TRUST IN HRI

Hancock and collaborators have defined human-robot trust as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others" [5]. In other words, one can trust a robot if its actions support both its own and the person's intentions. Nevertheless, in order to understand the development of this construct, Hancock et al. have reviewed factors affecting human-robot trust using a meta-analytic method and have identified three main elements: human-related, robot-related and environmental [6]. Human-related factors may include, for instance, prior experience, situation awareness and personality traits. While robot-related factors, which have the strongest currently known influence on trust, involve characteristics such as reliability, transparency or proximity. Finally, the environmental aspects accommodate task-related attributes and team collaboration elements as in-group membership or shared mental models. Later, based on this review of factors, Schaefer has developed a scale for the measurement of human-robot trust covering this triadic model of human, robot and environmental-related elements [17]. In relation to our work, the act of justifying a fault is associated to the transparency factor, which is robot-related.

4 THE ROBOTIC PEER THAT PLAYS TANGRAM

For the purpose of investigating the impact of recovery strategies during a collaborative task, we chose to use a scenario based on the Tangram game. This is a puzzle game that consists of putting together seven shapes, named "tans", in order to form more complex shapes (see Figure 2). Participants interact with a NAO robot, taking turns to solve the puzzle collaboratively. One of the reasons why this particular game was chosen is due to its simplicity, based on the advice of avoiding difficult and demanding tasks mentioned by Sarkar et al. [16].

During the task, the decisions and behaviours of the NAO robot are fully autonomous. This was achieved by relying on a Tangram application previously developed [1], which was built on top of SERA, a development ecosystem that merges techniques from computer animation and social agents to allow for the creation and animation of social robots [12]. The architecture of the resulting system is illustrated in Figure 1. As shown, participants interact using only a touchscreen interface and play the Tangram game with the social robot. The game logic, the decisions of the robot and the behaviour planner are divided into three separate components that communicate with each other using a middleware system [11], which are the Game Application, the Decision Maker and the Skene, respectively. A final component, NAOThalamus, is used to interface with the robot itself, through the NAOqi framework, which includes a text-to-speech component, a gaze module and all the animations the robot can perform.

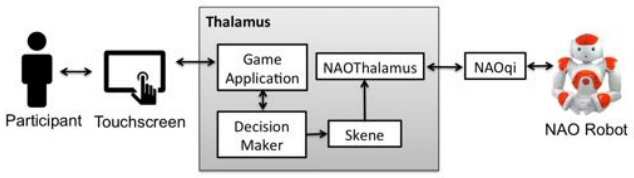

Figure 1: The system architecture.

As previously mentioned, the robot and the participant play the game collaboratively, where each one is allowed to move one piece per turn. The robot's verbal and non-verbal behaviour is

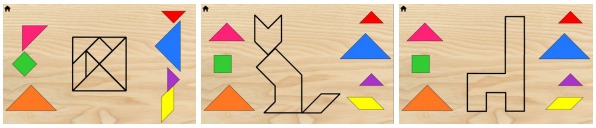

Figure 2: Set of Tangram puzzles for the user study.

essentially focused on reflecting the progress made by the user. The communication is asymmetric and in Portuguese. The robot always starts with an utterance referring its turn and it randomly chooses a piece to place in the puzzle. Then, it announces to the participant his/her turn during which, if he/she has some trouble rotating or placing the piece, the robot provides some advice or hints. The robot also gives compliments after a move made by the participant. To make the robot seem less predictable and artificial, there is a random chance associated to the performance of each behaviour. In the end of each puzzle, the robot celebrates with a joyful utterance and gesture, and also announces the number of puzzles that are left to finish the task.

Regarding the gaze of the robot, it always looks at the participant while talking to him/her and looks at the touchscreen otherwise. Moreover, the robot has no memory of previous moves or events. It only knows the name of the participant in order to establish rapport. This is introduced by the experimenter in the beginning and the robot uses it in the middle of some utterances.

Finally, the puzzles chosen were the square, the cat and the chair, in this order (Figure 2). The first puzzle was easy as the pieces were already with the correct orientation and their final positions were highlighted. For the second and third puzzles, only the outline of the final shaped was highlighted and the pieces were not with the correct orientation, making it harder to complete.

5 USER STUDY

Using the previously described scenario, a study was conducted to analyse the impact of a technical failure by a social robot during a cooperative setting.

5.1 Hypothesis

By trying to address how a technical failure affects the perception of trust in the robot, we posed two hypotheses for the experiment, as follows:

- **Hypothesis 1** - A technical failure of a social robot in a cooperative task will have a negative effect on the trust towards the robot.
- **Hypothesis 2** - A social robot that reveals transparency by justifying a technical failure during a cooperative task will mitigate the negative effect on the trust towards it.

5.2 Experimental Design

A between-subjects design was used in our user study, in which participants had to play the Tangram game with a stationary NAO robot (see Figure 3). We manipulated two variables: the recovery strategy (justifying or not) and the failure consequence (restart or continue). By justifying the fault, we aimed at associating the recovery strategy to one of the robot-related factors that influence trust, the transparency. By increasing the consequence of the failure on the task, we aimed at creating a situation where the human is directly penalised, since the task was collaborative. Furthermore, this manipulation is inspired by the fact that most real error situations during user studies require the task to start again. Additionally, there was a control group in which the robot did not fail at all. Therefore, the study had a total of 5 conditions:

- **Control Condition** - The robot did not have a failure.
- **Justification Strategy & The Task Continues** - After the failure, the robot attributes it to a technical problem. The game continues from the same moment it stopped and the participant can finish the task.
- **Justification Strategy & The Task Restarts** - After the failure, the robot attributes it to a technical problem. The progress of the game is lost and it restarts so that the participant has to play from the beginning.
- **No Recovery Strategy & The Task Continues** - After the failure, the robot says nothing. The game continues from the same moment it stopped and the participant can finish the task.
- **No Recovery Strategy & The Task Restarts** - After the failure, the robot says nothing. The progress of the game is lost and it restarts so that the participant has to play from the beginning.

For the failure conditions, the simulation of the error occurred in the middle of the third puzzle. While mentioning its turn, the robot stutters in the middle of a sentence saying "It's myyyyyyyyyyy" and then freezes for 50 seconds. After that, the game application presents one of two possible responses: continue or restart – corresponding to the manipulation of the failure consequence. In each of these groups, one of two possible recovery strategies can be presented – the robot justifies the failure by saying the sentence "There was a failure in my speech module. Let's continue/restart" or it says nothing.

Regarding the two conditions with the more severe failure consequence, i.e. when the robot needs to restart, participants played the same sequence of puzzles in the same order. After restarting, although the robot randomly selects the pieces and each puzzle may have a different order of moves, the remaining behaviours occur in a similar manner.

5.3 Experimental Procedure

Participation in the study was individual. The experiment was divided in 3 phases. The first consisted of having participants filling in the initial questionnaire related to their expectations of the robot before they interacted with it. The second experimental phase was to play a set of three Tangram puzzle games with the NAO robot on a touchscreen (see Figure 3). Finally, in the last phase, participants repeated the questionnaire they did at the start. Participants were also informed that they would stay alone in the room and were expected to leave at the end of the experiment or in case they wanted to interrupt the experiment.

5.4 Dependent Measures

Two dependent measures were used on the data analysis:

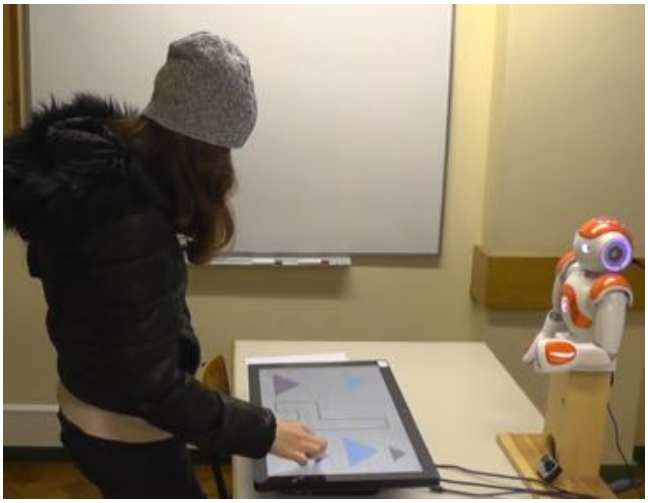

Figure 3: Participant playing a Tangram game with NAO robot on the touchscreen.

- **Trust** was accessed with the 14-items subscale of the Human-Robot Trust Questionnaire [17] in two time points, before and after the interaction with the robot. The assessment of the trust before the interaction was used as a covariate to measure the effect of the failure. The decision to use the 14-items subscale is due to the fact that it is focused on the functional capabilities of the robot.
- **Impact of the failure on the task** was accessed on a single item question ("Identify the impact of the failure on the task") with a Likert scale ranging from 1 ("Not severe") to 5 ("Very much severe").

5.5 Participation

The study was conducted at a Portuguese university and there was a total of 107 participants in the experiment. Five participants were excluded due to an unexpected technical failure in the system during the interaction with the robot. One more participant was excluded given that he left the room to ask for support immediately after the robot's failure, leading him not to hear the recovery of the robot. We also identified and excluded four outliers regarding the trust levels using a step of 1.5 x IQR (interquartile range), leaving us a total of 97 participants (71 males and 26 females) with ages ranging from 17 to 41 years old ($M = 22.26 \pm 4.51$). We tried to balance around 20 per condition: 16 in the control condition, 38 in the group where the task continues (18 in the Justification Strategy condition; 20 in the No Recovery Strategy) and 43 in the group where the tasks restarts (21 with Justification Strategy; 22 with No Recovery Strategy).

6 RESULTS

Initially, we conducted a reliability analysis (Cronbach's α) to assess the internal consistency of the 14-items subscale by Schaefer [17]. Since the reliability was too low ($\alpha = 0.46$), we excluded the two most inconsistent items, which were "Dependable" and "Predictable". Without them, the internal consistency became acceptable ($\alpha = 0.72$). Therefore, the following results reporting trust levels refer to this 12-items subscale. Furthermore, a normality analysis was conducted, which revealed that the dependent variable of trust did not follow a normal distribution (Shapiro-Wilk test). We also verified the homogeneity of variances assumption ANOVA grounds on with the Levene's test and it was not significant (p=0.998).

6.1 Manipulation Check

By applying the Mann-Whitney test, we observed that the manipulation of "restarting" caused a significant difference (Figure 4) on the impact on the task caused by the failure ($U = 394; p = 0.001; r = 0.38$). The failure was perceived to have less impact in the group where the game continues after the robot's failure ($M = 1.90 \pm 0.72$) when compared to the group where the game restarts ($M = 2.74 \pm 1.11$).

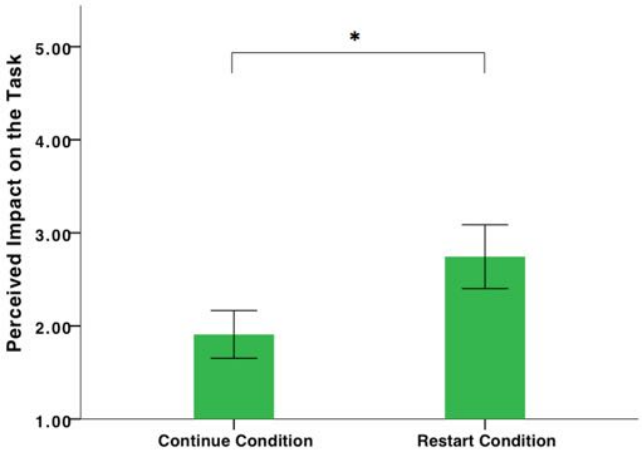

Figure 4: Differences in the perceived impact on the task caused by the failure when the game continues and restarts. (*p = 0.001)

6.2 Effect of the Failure

A 1-way ANOVA test was applied to analyse the overall effect of the robot failing. We compared the participants' trust towards the robot when it failed and when it did not, controlling for the trust levels reported before the interaction. Results showed a significant difference ($F = 12.97; p = 0.001$). As shown in Figure 5, participants in the group where the robot did not fail showed higher trust levels towards the robot ($M = 90.94 \pm 7.93$) than the group where the robot failed ($M = 84.28 \pm 9.60$).

6.3 Effect of Recovery Strategy

To analyse the impact on the trust of the recovery strategy for the conditions where the robot fails, we used a Factorial ANOVA test after applying a rank transformation to the data. There was no significant main effect for the recovery strategy ($F = 1.25; p = 0.268$) nor for the failure consequence ($F = 1.14; p = 0.288$). However, we found a significant interaction effect between these two variables ($F = 4.17; p = 0.045$).

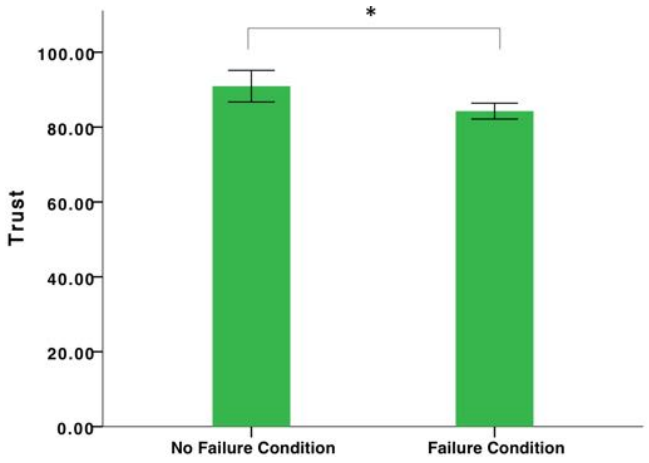

Figure 5: Difference in the trust levels towards the robot when there was a failure and when there was not. (*p = 0.001)

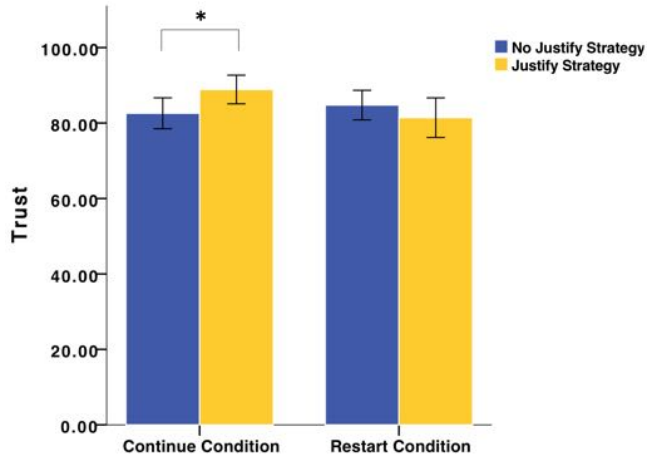

Figure 6: The effect of the recovery strategy at individual levels of the failure consequence. (*p = 0.033)

Consequently, we conducted a simple effects analysis by looking at the effect of the recovery strategy at individual levels of the failure consequence (Figure 6), using a Mann-Whitney test. For the group with the less severe failure where the task continues, the trust levels were significantly different between the recovery strategies ($U = 107; p = 0.033; r = 0.35$). More precisely, in the case where the robot justified its failure, the trust was significantly higher ($M = 88.89 \pm 7.63$) than in the case where it did not justify ($M = 82.58 \pm 8.77$). However, for the group with the most severe failure where the task restarts, there was no statistically significant difference ($U = 201.5; p = 0.473$) on the trust between the justification recovery strategy ($M = 81.43 \pm 11.54$) and the no recovery strategy condition ($M = 84.76 \pm 8.83$).

6.4 Failure Mitigation

An additional statistical analysis compared the trust levels in the control condition and each of the failure conditions, using a Mann-Whitney test. The trust levels were significantly different between the control condition and: (1) the no recovery strategy where the task continues ($U = 77; p = 0.007$); (2) the justification strategy where the task restarts ($U = 88.5; p = 0.013$); (3) the no recovery strategy where the task restarts ($U = 108; p = 0.045$). However, the difference between the trust levels in the control condition and the justification strategy where the task continues was not statistically significant ($U = 119.5; p = 0.403$).

7 DISCUSSION

Our results support **Hypothesis 1**, reflecting the negative impact of faulty robot's performance in the trust level towards it.

Moreover, we extend the findings of Lee et al. [7] and Salem et al. [15] by analysing a different type of failure and recovery strategy during a cooperative game, where the robot has the role of a peer.

According to **Hypothesis 2**, we expected the recovery strategy of justification to reveal transparency from the robot and consequently mitigate the negative impact of its technical failure. This was partially confirmed by the interaction effect of both the recovery strategy and the task impact on the trust levels. Additionally, the trust levels were significantly different between the control and each one of the other conditions, except for the justification strategy where the task continues. This recovery strategy was able to mitigate the negative impact on the trust levels when the failure's consequence was less severe by continuing the game. On the contrary, when the failure's consequence was more severe and participants had to restart the task from the beginning, the recovery strategy of justifying the failure was not able to mitigate its negative impact on trust.

We believe the recovery strategy of justifying the failure was weak for the case of restarting the task. Especially due to the cooperativeness of our setting, in which this consequence affected the participants' progress in the task. Also, other findings from social psychology, relating to the specific recovery strategy of apologising, revealed that extensive apologies are required to mitigate more severe harms [9].

8 CONCLUSION AND FUTURE WORK

Our work shows the negative impact of a technical failure on trust in a collaborative setting and how the fault justification as a recovery strategy can mitigate this negative impact.

Our main contribution lies on a user study with an in-person interaction with a fully autonomous robot that executes a recovery from a technical failure. This is particularly important in applications with autonomous robots, which should also be able to recover from failures autonomously. We investigated the effect of two different consequences for the failure, continuing or restarting the task, which is inspired in real-world situations where, almost inevitably, there are errors that require the task to restart. Furthermore, we investigated the effect of a recovery strategy that was not yet explored in human-robot trust research, where the robot justifies the failure (or ignores it) and is therefore associated to a more (or less) transparent agent. Due to the fact that the robot simulates

awareness of the failure's cause and can consequently simulate an autonomous recovery, such recovery strategy is perceived as natural.

Our results have shown that faulty behaviour by a social robot during a cooperative task, such as a puzzle game, is indeed perceived as less trustworthy. However, our main result indicates that justifying the failure as a recovery strategy can mitigate its negative impact on the trust, but only when the consequence of the failure is less severe. On the other hand, when the failure is more severe, the recovery strategy has no effect on the trust. This might happen because of a higher expectation from the participants for the recovery strategy, where only a justification may be perceived as inconvenient.

We believe our results can be generalised for similar low-risk and cooperative situations, such as conversational, entertainment or companion robots. Every situation that compromises the trust towards a social robot must be addressed, which according to the human-robot trust definition, happens when there is a discrepancy between the robot's perceived intention and its actions. Therefore, these findings can be a valuable lesson when developing robots in HRI scenarios, and we hope they can contribute to the development of more reliable and trustworthy robots, which reveal recovery strategies capable of mitigating the negative effect of possible failures. More importantly, our results also point to the fact that mitigation strategies should be tailored according to different factors, such as task type, failure type and failure severity.

However, being able to detect that there was an error is still tricky. Only in the case where the system can detect that there was some fault, it makes sense that the robot can recover and apply the correct strategy. Another point worth mentioning is that some participants might have perceived the justification of the robot as a mere acknowledgement, which could have been disambiguated by a manipulation check in the questionnaire.

As future work, we plan to improve our study and develop a robot using adaptive models where it changes its recovery strategy accordingly to the failure severity. In terms of type of failure, we also aim to explore the use of task-related failures instead of only technical. It would also be interesting to explore what the effect is regarding the trust of using recovery strategies that include some emotional content.

ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, through the project AMIGOS (PTDC/EEISII/7174/2014), the project RAGE (Ref. H2020-ICT-2014-1/644187), and the project LAW TRAIN (Ref. H2020-FCT-2014/653587). Filipa Correia and Carla Guerra acknowledge their FCT grants (Ref. SFRH/BD/118031/2016 and SFRH/BD/118006/2016, respectively).

REFERENCES

[1] Beatriz Bernardo, Patrícia Alves-Oliveira, Maria Graça Santos, Francisco S Melo, and Ana Paiva. 2016. An Interactive Tangram Game for Children with Autism. In *International Conference on Intelligent Virtual Agents*. Springer, 500–504.

[2] Daniel J Brooks, Momotaz Begum, and Holly A Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 487–492.

[3] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology* 6 (2015).

[4] Adriana Hamacher, Nadia Bianchi-Berthouze, Anthony G Pipe, and Kerstin Eder. 2016. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 493–500.

[5] PA Hancock, DR Billings, and KE Schaefer. 2011. Can you trust your robot? *Ergonomics in Design* 19, 3 (2011), 24–29.

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

[7] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 203–210.

[8] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[9] Ken-ichi Ohbuchi, Masuyo Kameda, and Nariyuki Agarie. 1989. Apology as aggression control: its role in mediating appraisal of and response to harm. *Journal of personality and social psychology* 56, 2 (1989), 219.

[10] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 501–506.

[11] Tiago Ribeiro, André Pereira, Eugenio Di Tullio, Patrıcia Alves-Oliveira, and Ana Paiva. 2014. From Thalamus to Skene: High-level behaviour planning and managing for mixed-reality characters. In *Proceedings of the IVA 2014 Workshop on Architectures and Standards for IVAs*.

[12] Tiago Ribeiro, André Pereira, Eugenio Di Tullio, and Ana Paiva. 2016. The SERA ecosystem: Socially Expressive Robotics Architecture for Autonomous Human-Robot Interaction. In *Enabling Computing Research in Socially Intelligent Human-Robot Interaction: A CommunityDriven Modular Research Platform*.

[13] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 101–108.

[14] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

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

[16] Satragni Sarkar, Dejanira Araiza-Illan, and Kerstin Eder. 2017. Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker. *arXiv preprint arXiv:1703.02335* (2017).

[17] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).