Robot Broken Promise? Repair strategies for mitigating loss of trust for repeated failures

Robot Broken Promise? Repair strategies for mitigating loss of trust for repeated failures

Birthe Nesset[1], Marta Romeo[1], Gnanathusharan Rajendran[2] and Helen Hastie[1]

***Abstract*— Trust repair strategies are an important part of human-robot interaction. In this study, we investigate how repeated failures impact users' trust and how we might mitigate them. Specifically, we look at different repair strategies in the form of apologies, with additional features to them such as warnings and promises. Through an online study, we explore these repair strategies for repeated failures in the form of robot incongruence, where there is a mismatch of verbal and non-verbal information given by the robot. Our results show that such incongruent robot behaviour has a significant overall negative impact on participants' trust. We found that the robot making a promise, and then breaking it, results in a significant decrease in participants' trust, when compared to a general apology as a repair strategy. These findings contribute to the research on trust repair strategies and, additionally, shed light on how robot failures, in the form of incongruences, impact participants' trust.**

I. Introduction

With the growing development of social robots, various robot failures need to be mitigated for. With social robots, not only should one handle technical shortcomings and errors from the robot, but a whole new spectrum of interaction failures need to be taken into consideration as well. The robot might say or do something unexpected or uncalled for, which breaks with the users' understanding of the interaction or their perception of the robot's abilities [1], [2].

When this happens, different trust repair strategies could be implemented to mitigate some of the loss of trust [3]. Usually, the robot will apologise for committing a failure, often using additional tools such as emotions or promises [4], [3], [5]. Apologies seem to work well in single-failure situations. However, once a failure has happened, this does not necessarily mean that the interaction stops nor that additional failures will not happen. In fact, if the failure does not have any severe or harmful consequences, users might continue with the interaction, thus potentially causing the same failure to repeat itself [6].

An example of unexpected mannerisms is a robot's incongruence. The robot might respond to a user with emotions or facial expressions that do not fit the conversation, or it might behave in an unanticipated manner, such as turning at unexpected times or answering a question while addressing the incorrect person [7], [8]. These incongruencies, though not always wrong nor harmful, could be perceived as a robot failure by the user, causing potential decreases in trust and negatively impacting the users' attitude towards the robot's capabilities [9]. In this paper, we aim to expand on the current literature on repair strategies in the form of apologies, by adding extra features to them such as warnings and promises, in interactions with a repeated failure in the form of a robot incongruence between the robot's gestures and its verbal command.

This study can be divided into two main research questions. In the first part, we look at how introducing a robot failure might impact users' trust. With this in mind, we propose the following hypothesis: **H1**: A robot failure will significantly reduce users' trust.

The second part of the study looks at the implementation of different repair strategies and how they might influence users' trust when a repeated failure takes place. We introduce four different repair strategies: apology, apology with a promise, apology with a warning and no apology. We hypothesise the following **H2**: The implementation of an apology with a promise will lead to a more extensive decrease in the users' trust when a repeated failure is introduced, as compared to other forms of repair strategies.

Additionally, we investigate what communication method people follow when faced with an incongruence and forced to pick between the two. We also investigate if there are any individual differences (such as negative attitudes towards robots or propensity to trust) among the participants, based on whether they prefer verbal or non-verbal instructions, and if they have any reasoning behind their preference.

II. Related Work

Trust is stated to be an important aspect of any HRI research when it comes to creating acceptance amongst users, especially in scenarios of collaboration where it is necessary that the users follow robot-produced suggestions and decisions [10], [11]. A commonly used definition in Human-Robot Interaction (HRI) focuses on human trust in automation and can be defined as the attitude that: "*an agent will help achieve an individual's goal in a situation characterised by uncertainty and vulnerability*" [12]. If the robot's behaviour deviates from the trusted behaviour, for instance by committing a failure, this can impact the users' trust [13].

A robot failure has been defined as: "*a degraded state of ability which causes the behaviour or service being performed by the system to deviate from the ideal, normal or correct functionality*" [14]. Even though robot error is generally regarded as a performance-based factor, the user

This work was funded and supported by the UKRI Node on Trust (EP/V026682/1) https://trust.tas.ac.uk

[1]School of Mathematical and Computer Sciences, Heriot-Watt University, UK. bn25@hw.ac.uk, m.romeo@he.ac.uk, h.hastie@hw.ac.uk [2]School of Social Sciences, Heriot-Watt University, UK. t.rajendran@hw.ac.uk

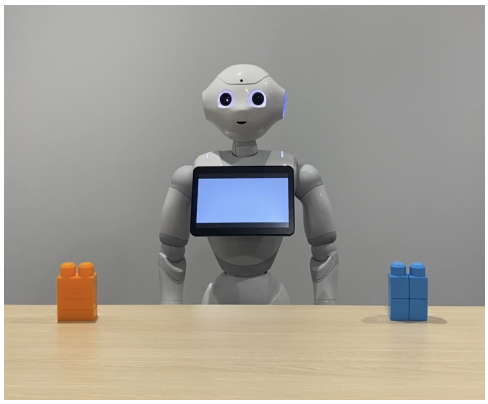

Fig. 1. Setup for block selection with Pepper. The colours of the blocks vary between green, orange, blue and purple. The robot would always start in the position shown in this figure, before moving towards one of the blocks. When the robot moves it simultaneously gives the verbal indication and the corresponding non-verbal gestures and the colour changed on the tablet.

and their environment will also impact how the error is interpreted and to what extent this impacts trust, for example errors committed in stressful or novel environments [15]. Giuliani et al. [2] classified failures according to their type: Technical failures - caused by technical shortcomings in the robot and social norm violations - when the robot deviates from the social script or uses inappropriate social signals. A robot incongruence can often be perceived as a social norm violation, although it might derive from a technical shortcoming in the robot as well [2]. Previous work has looked at robot incongruence in the form of facial expressions [16], congruent and incongruent approach behaviour when giving different information [17], incongruent robot gestures [9] and the use of different emotional cues [18], [7]. While none of these studies looked specifically at the relation between robot incongruence and users' trust, Tsiourti et al. [7] found that when an observer receives incongruous emotional information across the auditory and visual modalities, the incongruences confused their users and impacted the robot's likability and believability. Additionally, Salem et al. [9] found that incongruence between the robot's gestures and what it said had a significant impact on users' task performance, the perceived anthropomorphism of the robot and the mental model users had of the robot's abilities.

When robot failures take place, different forms of repair strategies can be implemented to mitigate some of the potential losses in trust [19]. As discussed by Esterwood and Robert [19], the human-human literature has classified these efforts as either apologies, denials, explanations, or promises. The HRI literature on which is the best repair strategies is rather mixed. At times. promises have been shown to be a more effective repair strategy than general apologies or denials [4], on the other hand explanations have been shown to work better than apologies in certain instances, when an apology is not necessarily warranted [20], [21]. Recent work [22] found no real difference between repair strategies (apology, explanation, promises), with only denials behaving significantly worse.

Additionally, previous work has also looked at embellishing established repair strategies. For example, by having the robot display remorse [23] while apologising, or by acknowledging the mistake and handling it [21]. Though a lot is currently undetermined in regards to HRI repair strategies, there seems to be a general consensus on the efficacy of apologies and acknowledging that mistakes have taken place. We, therefore, design our repair strategies with this in mind, creating a general apology as a baseline, and then adding additional features in the form of a promise and a warning.

III. Method

In our video-based survey study, participants were tasked with choosing between two blocks of different colours, based on the robot's verbal and non-verbal indications (see Figure 1). This study was conducted online, via Microsoft Forms Survey. Participants were recruited through Prolific[1] and randomly allocated to one of four conditions, corresponding to the repair strategy the robot implemented after its failure. Through Prolific, multiple screening criteria were implemented to ensure that no vulnerable participants, or those suffering from any form of colour vision impairment, were invited to participate in the study.

Participants were debriefed and asked for consent before they started the study. We gathered data on their age, gender, propensity to trust robots [24] and their negative attitudes towards robots [25]. They were then again debriefed on the task they were to complete and asked to identify four different colours. These colours were later used for the blocks in the study. This test was included as an additional screening process to limit the study's exposure to people suffering from colour blindness, beyond the preliminary screening implemented in Prolific.

The main part of the study was divided into three different parts, see Figure 2 for a flow diagram of the study. To ensure that the failures had the same effect on all participants, the ordering of these parts was kept the same. In all three parts, participants are tasked with selecting a coloured block with the help of a Pepper robot, see Figure 1. The blocks were placed on a table and were either green, orange, blue or purple. The robot indicated a block by saying the name of the colour out-loud, while simultaneously indicating the block by moving towards it, gesturing to it with both hands and a head tilt, and changing the colour of its tablet to the same colour as that of the block.

For the first part of the task, participants watched three of these videos and selected a colour after each video. After the third video, they were also asked to complete the Trust in Automated Systems (TAS) scale [26], we will refer to this measure as TAS1 throughout the rest of the paper. In the first part, no robot failure is introduced, meaning that there was no difference between what the robot said and what it gestured towards. In the second part, participants watched again three videos. However, in the third video of

[1]https://www.prolific.co

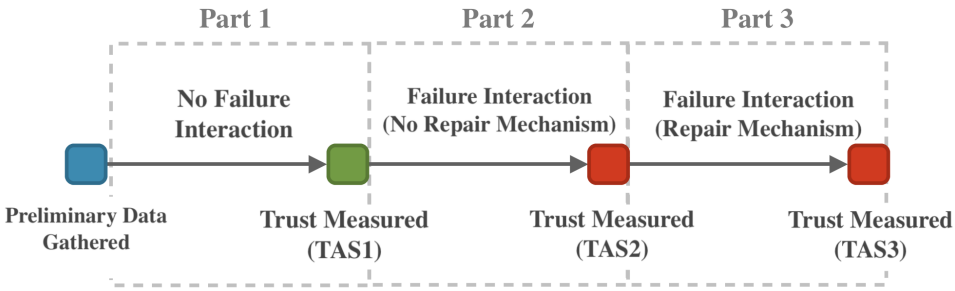

Fig. 2. Flow diagram of the three parts of the study. The trust measured in the first part of the study, before any failure takes place is referred to as TAS1. TAS2 is the trust measured after the first failure with no repair strategy, and TAS3 is the trust measured after the second failure with the repair strategy implemented. Each data gathering point is marked by a box on the figure, where blue is the preliminary data, green represents the no failure data and red is the failures data.

TABLE I

PAIRED SAMPLES T-TEST LOOKING AT THE DECREASE OF TRUST WITH EACH ROBOT FAILURE

Measure 1		Measure 2	Test	Statistic	z	df	*p*	Effect Size
TAS1:	-	TAS2:	Student	22.597		184	$< .001$	1.661
			Wilcoxon	16863.500	11.547		$< .001$	0.982
TAS2:	-	TAS3:	Student	3.227		184	0.001	0.237
			Wilcoxon	8841.000	3.917		$< .001$	0.356

the second part a robot failure is introduced, where the robot gestures towards one colour while saying the other. After this interaction, participants were tasked with selecting one colour, they could either choose the one the robot said or the one it gestured towards.

They were then again asked to complete the trust questionnaire for a second time (TAS2). For the last part of the task, participants were given one out of four repair strategies. This could either be a general apology, an apology with a promise, an apology with a warning or no apology at all. The general apology was: "*I am sorry I seemed to have made a mistake*". The apology with a promise was: "*I am sorry I seem to have made a mistake, I promise this will not happen again*", and the apology with a warning was: "*I am sorry I seem to have made a mistake. I have not been trained on this, so please be aware that it might happen again*". When no apology was implemented, participants went on to the next part of the study without the failure being acknowledged.

After the repair strategy was observed, participants completed the third and final part, where they again watched three videos. Similarly to the second part, the first two videos did not portray any failures, while the final one included another robot failure. After this, participants completed a third and final trust questionnaire (TAS3).

The voice and gestures indications remained the same throughout all nine videos. The robot was always saying "*Please select the xx block*" (where xx refers to one of the possible colours), while the non-verbal signal was the robot moving to centre itself in front of a block, indicating the block with both hands, tilting its head towards the block and changing the colour on its tablet to match the colour of the block in front of it.

At the very end of the study, participants were asked to express whether they followed the voice, gestures or both indications given by the robot. They also had the option to provide a reason (if they had any) behind their choice in the failing interactions.

To ensure that participants watched all the videos to the end, multiple attention checks were set in place. After each video, a number is displayed that participants are asked to provide after the video ends. In the apology video, this number is said out loud by the robot at the end. We also implemented an attention check after the second trust scale (TAS2) asking participants to select agree on a five-point Likert scale. All these measures were set in place to ensure that the participants watched every video carefully and answered the questions to the best of their ability. Additionally, we added one manipulation check at the end of each part, asking if the robot made any errors in the video. Participants who failed to correctly detect when the robot made a mistake were removed from the final data set, together with those who failed any of the attention checks.

IV. RESULTS

A total of 200 participants were recruited to participate in the study, with about 50 participants randomly allocated to each condition. After excluding participants who failed the attention checks, the colour test, or the manipulation test, 185 participants remained. We recruited 86 participants who self-identified as females, 98 males and one transgender female. The mean age was 37.1 years with an SD = 11.75.

A. Repeated Robot Failure

To test our first hypothesis: "*A robot failure will significantly reduce users' trust*", we conducted a paired samples t-test to determine if the introduction of a robot failure had a significant impact on participants' trust, see Table I. We compared the trust measured by the TAS-scale in

TABLE II

RESULTS OF ANOVA COMPARING THE DECLINE IN TRUST BETWEEN THE FOUR REPAIR STRATEGIES.

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Condition:	5.476	3	1.825	3.732	0.012	0.058
Residuals	88.527	181	0.489			

the non-failure interaction with the trust measured after the first failure is introduced (TAS1 vs TAS2). The student t-test showed a significant difference between trust measured before (M = 5.22, SD = 0.89) and after (M = 3.46, SD = 0.91) the first failure with $t(184) = 22.597$, $p < .001$ and an effect size of 0.982. Due to a deviation in normality showed by Shapiro-Wilk's test, a Wilcoxon signed-rank test was used when comparing the trust after the first failure with the trust after the second failure (TAS2 vs TAS3). The Wilcoxon test showed a significant decrease in measured trust before (M = 3.46, SD = 0.91) and after (M = 3.29, SD = 0.99) the second failure with Z = 8814, $p < .001$, and an effect size = 0.356. Both effect sizes are reported as the matched rank biserial correlation for comparison. See Figure 3 for a line diagram of the mean trust and trust decline for TAS1, TAS2 and TAS3.

For the second hypothesis: "*The implementation of an apology with a promise will lead to a more extensive decrease in the user's trust when a repeated failure is introduced, as compared to other forms of repair strategies*", we investigated how the implementation of different repair strategies impacted participants' trust when failures were repeated. We compared our four different repair strategies: a general apology, an apology with a promise, an apology followed by a warning, and no apology at all.

We compared these conditions using an ANOVA, see Table II, and found that there was a significant difference between the decline in trust depending on which repair strategy the participant was exposed to. The measure 'decline in trust' is the difference between the trust after the first failure compared to the trust after the second failure, where the different repair strategies had been enacted. The ANOVA gave us the following results $F(3,181) = 3.73$, $p = 0.012$ and an estimated effect size $\eta_p^2 = 0.058$. Furthermore, a post-hoc evaluation, looking at the different relationships between the repair strategies, found a significant difference between the general apology and the apology with a promise, see Table III. The mean difference between the two repair strategies was 0.466 with a SE = 0.146, the p_{tukey} = 0.009 and the effect size was 0.667 (Cohen's d). The mean for each repair strategy individually can be found in Table IV and a box-plot of each repair condition is presented in Figure 4.

B. Participants' Communication Preference

In addition to verifying our main hypothesis, we also wanted to have a closer look at the participants' communication preferences when faced with a robot failure.

Looking at participants' block selection after the first failure took place, 25.41% of the participants based their decision on the robot's gestures, while 74.59% on the verbal

TABLE III

POST HOC COMPARISONS BETWEEN THE REPAIR STRATEGIES.

		Mean Diff	Cohen's d	p_{tukey}
Apology:	NoApology:	0.194	0.278	0.552
	Promise:	0.466	0.667	0.009
	Warning	0.115	0.164	0.858
NoApology:	Promise:	0.272	0.389	0.248
	Warning	−0.079	−0.114	0.947
Promise:	Warning	−0.351	−0.502	0.072

TABLE IV

TABLE OF DESCRIPTIVE STATISTICS FOR THE MEAN DECLINE OF TRUST BETWEEN TAS2 AND TAS3 FOR EACH REPAIR STRATEGY.

Repair Strategy	Mean	SD	N
Apology	0.026	0.614	45
Warning	-0.089	0.670	48
No apology	-0.168	0.713	45
Promise	-0.440	0.786	47

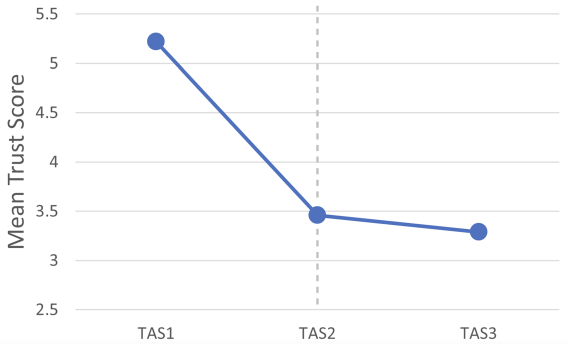

Fig. 3. Line diagram of the mean trust measured in TAS1, TAS2 and TAS3. The decline in trust was significant both between TAS1 & TAS2, and TAS2 & TAS3.

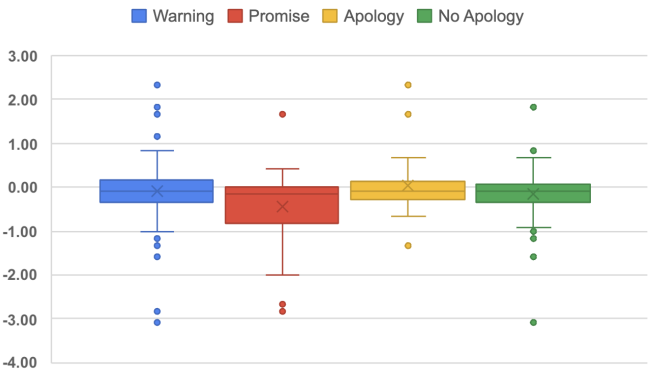

Fig. 4. Box-plot of decline in trust between TAS2 and TAS3, for each repair strategies when failure is repeated.

TABLE V

INDEPENDENT SAMPLES T-TEST LOOKING AT THE PARTICIPANTS' COMMUNICATION PREFERENCE COMPARED TO THEIR PRELIMINARY PERSONAL TRAITS.

	Test	Statistic	df	p	Effect Size
NARS:	Student	−2.124	183	0.035	−0.359
	Mann-Whitney	2645.500		0.059	−0.184
P2T:	Student	1.318	183	0.189	0.223
	Mann-Whitney	3653.000		0.194	0.126
Age:	Student	−2.436	183	0.016	−0.411
	Mann-Whitney	2416.500		0.009	−0.255
Gender:	Student	−1.003	182[2]	0.317	−0.169
	Mann-Whitney	2946.500		0.317	−0.085

[1]Participants split into those who self-identified as male and female. However, one participant was not included in this particular analysis because there were not enough members of this category to make any statistical meaningful comparisons.

command. When the failure was repeated, 6.49% of the participants decided to switch the modality they were following, e.g from the voice for the first failure to gestures for the second or visa-versa.

To investigate participants' preferences further, we had a closer look at the preliminary personal traits gathered: age, gender, their propensity to trust robots and negative attitudes towards robots. This was done by using a quasi-experiment, that is participants were allocated to one of two groups based on their communication preference during the first failure. Using an independent sample t-test, we found a significant difference between the two participant groups (gesture vs voice), in terms of negative attitudes towards robots ($t(183)$= -2.124, with p = 0.035, and an effect size of -0.359). The participants who favoured voice had a significantly higher negative attitude towards robots (M = 2.99, SD = 0.55) than the participants who preferred gesture (M = 2.79, SD = 0.58). Due to a deviation from normality, we used a Mann-Whitney U test to look at the parameter for age and found a significant difference between the two participants groups (U = 2417, p = 0.009 and an effect size of -0.255, given by the rank biserial correlation). Our results showed that participants who followed the robot's voice were older (M = 38.3, SD = 12.07) than the participants who followed the robot's gestures (M = 33.5, SD = 10.07). Participants' preferences had no relation to either gender or propensity to trust robots, see Table V.

Finally, we asked participants an optional open-ended question to glean some insight into their motivation behind choosing to follow the voice of the robot or its gestures. First, we asked them what communication method they followed when faced with the robot failure, either gesture, voice or both. We then asked if they could elaborate on why they followed one over the other. This question was optional and participants could choose to not respond. Comments from participants, who were not able to identify what they had previously chosen have not been included in the final qualitative data analysis. However, the same participants have not been excluded from the overall study as correct recall of their choice was not requested during the study itself. A total of 16 participants were unable to correctly recall their selection preference, 120 decided to follow the voice (V), 41 the gestures (G) and 8 indicated both voice and gestures (B).

A thematic analysis was then conducted by two researchers to code any underlying themes in the answers given by the participants. A total of 36 participants (25V, 6G and 5B) gave no reasoning behind their selection. 22 participants said they followed the chosen modality because it seemed more reliable (16V and 6G). 21 participants reasoned that they made their decision based on the additional information given by the robot tablet (2V, 17G, 2B). 15 participants argued they made their decision due to a perceived superiority of the communication medium (13V and 2G), while 11 did so because they found the selected modality more trustworthy (6V, 4G and 1B). Interestingly, 22 participants reported that they decided to follow the voice because they viewed it as a separate entity from the physical body of the robot, while 36 said they did so because they were following the robot's orders. Similarly, 8 participants following the robot's gestures said they did so because they followed the robot's decision.

V. DISCUSSION

In this study, we investigated how the implementation of a repeated robot failure, in the form of an incongruence, impacts users' trust, and how different repair strategies might mitigate its effect. We also looked at participants' preferred communication method when forced to choose to follow either the robot's verbal or non-verbal indication.

As presented in Section IV-A, the robot failure led to a significant decrease in trust between the first (no failure) interaction and the second (failure without a repair strategy) interaction (across all conditions). Furthermore, we also found a significant decrease in the participants' trust between the second interaction (failure without a repair strategy) and the third interaction (failure with a repair strategy). This supports our first hypothesis, that participants' trust will be impacted by the failure of the robot. Further, the second decline in trust, shows that this failure has a continuing impact on participants' trust. This is in spite of there being no consequences behind the participants decisions, by that we mean that no penalty or negative impact has been added as a result of the failure. This finding show how important it is to implement the right form of repair to help mitigate this trust loss, even for failures that do not have severe consequences.

With our second hypothesis, we aimed to look at how different repair strategies, in the form of apologies, impact participants' trust in repeated failures. The results presented in Section IV show a significant difference between the repair strategies deployed in terms of the magnitude in the 'decline in trust'. A post-hoc test show that in our case, implementing an apology with a promise will lead to a significantly larger decrease of the participants' trust, when compared to a general apology. This supports our second hypothesis: with repeated failures, implementing an apology with promise like the one used in this study will lead to a significantly larger decrease in trust. This could be due to the further impact broken promises have on participants' trust towards the robot [6], [5]. Implementing no apology or a warning

slightly mitigated the loss of trust compared to the promise, but was not significantly better. This is not to say that there will not be any significant impact over time, as these repair strategies were only implemented once. However, for this specific scenario, there will be a significantly smaller decline in users' trust if we use a general apology as compared to an apology with a promise that will later be broken.

In addition to testing our main hypothesis, we aimed to investigate how participants behaved once forced to choose to follow an incongruent behaviour from the robot. As presented in the results, we saw that the majority of the participants (around 75%) preferred to follow the robot' voice, while the rest went with its gestures. We also noticed that 6.5% of the participants changed their preferences. This shows that most participants were consistent in their preference. When asked to elaborate on their selection preferences, 79% of the participants seemed to have some form of justification or reasoning behind their selection preference, meaning that there is some underlying conscious reasoning behind their decisions.

We also found participants with different preferences, sometimes used the same argument when justifying their choice. When giving as explanation the reliability of the chosen medium we see that one participant went with gesture ("*I felt like the physical action of choosing a colour was stronger and less easy to get wrong than the spoken colour. I felt a mistake could more easily be made with the word than the actual action/gesture*") and one went with voice ("*I chose voice over what was seen on its screen as voicing something seems more intentional & less likely to be the mistake. Pointing to the wrong colour or displaying the wrong colour seems more of an accident than saying the wrong thing so assumed the speech was the correct intended answer*") because their selection was less likely to be a mistake.

In addition to this, an interesting trend emerged as 22 of our participants reasoned that they viewed the robot voice and its body as two separate entities ("*I listened to the comments and decided which colour do I need to choose since I thought robot not following instructions carefully*"), meaning that while the voice gave the order, the robot failed to follow its own orders (hence the incongruence). Again, voice was considered the superior means of communication ("*The voice seemed like the authority, and the images a visual guide to assist*") and the order to follow ("*I thought it was best to follow the verbal instruction*"). This hints to voice being a powerful element in human-robot interaction, able to shape the perceptions of our participants. We find this to be an interesting tendency that should be investigated further to see if it is an attitude that emerges with robot incongruence, or if it is due to the study being conducted online.

Furthermore, we also looked at the difference between participants who preferred gestures and the ones who went with the verbal indication in the first failure. Our results, as presented in Table V, showed that there was a tendency amongst the participants who preferred to follow the robots gesture, to have significantly lower negative attitudes towards robots and to be younger than the ones following the robot's voice. Though these results are interesting, they are in need of further investigation. However, they could potentially point towards some interesting opportunities to start creating behavioural measures for peoples' internal attitudes towards robots, something that could in the long term benefit the research in HRI, as an alternative to subjective questionnaires. Additionally, this could potentially improve tailoring of robot behaviour in correlation with the participants' attitudes.

However, this study is not without its limitations. There are certain limitations that follow using online sourced participants, and we urge to keep this in mind when implementing these findings in interactions that take place face-to-face. We also did not implement any consequences of the robot's failures, and though the failure still had an impact on the participants' trust, these results might vary when implemented in a situation different from the one in this study. Different repair strategies might be more suitable to mitigate trust loss depending on the severity and rapidness of the failure. We also purposefully designed the failure to have no consequence and limited the participants' information on the task to try and glean their natural inclination towards their communication preference without having them delve into it for too long. These results might therefore again vary when participants are more or less informed about the situation at hand. We find the results here to be an interesting start to the discussion of participants' communication preferences and is a useful contribution to the research on repair strategies in accordance with robot failures.

VI. Conclusion and Future Work

In this study, we investigated how a robot failure impacts participants' trust towards the robot, how different repair strategies in the form of apologies, impact participants' trust when faced with a robot that commits repeated failures, and we looked at participants' communication preferences when faced with failure in the form of a verbal/gesture incongruence.

Through an online evaluation, we implemented repeated robot failure in the form of a mismatch between what the robot was saying and what it was gesturing towards. After the first failure, we divided our participants into four different repair strategies (apology, apology with promise, apology with warning, and no apology). We then repeated the failure one more time. We investigated how this impacted participants' trust, and what communication method they preferred when forced to choose.

Our study results point towards a significant decrease in trust for each failure. We also looked at the impact of different repair strategies. Though previous work has shown some benefits of using promises when apologising [3], our results show that for situations, where the interaction is not likely to stop due to a failure, broken promises could worsen the trust between the participant and the robot when these failures are repeated. In these situations, it will therefore be more beneficial to use more general apologies to address the failure. We also found that the majority of the participants

preferred the voice command, though 25% went with the robot's non-verbal indication.

In future work, further investigation into peoples' communication preferences could lead to some useful insights and help to further the design of clear and transparent robots. Additionally, the interactions exhibited in this study show some interesting potential in the design of behavioural measures within HRI and peoples' internal attitudes. Finally, our work here is an interesting contribution to current research on repair strategies. These findings are worth investigating further. For instance, looking further into the impact of repeated failures over a more extensive time period could be of interest. Additionally, altering the failure type could change the impact of the repair strategies. In the future we might want to further look into how to optimise repair strategies in more vulnerable situations where the failures have consequences for the participants.

REFERENCES

[1] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[2] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations," *Frontiers in Psychology*, vol. 6, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00931

[3] C. Esterwood and L. Robert, "A literature review of trust repair in hri," *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 07 2022.

[4] P. Robinette, A. Howard, and A. Wagner, "Timing is key for robot trust repair," *Seventh International Conference on Social Robotics*, 10 2015.

[5] Y. Albayram, T. Jensen, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman, "Investigating the effects of (empty) promises on human-automation interaction and trust repair," *Proceedings of the 8th International Conference on Human-Agent Interaction*, p. 6–14, 2020. [Online]. Available: https://doi.org/10.1145/3406499.3415064

[6] A. Gneezy and N. Epley, "Worth keeping but not exceeding: Asymmetric consequences of breaking versus exceeding promises," *Social Psychological and Personality Science*, vol. 5, no. 7, pp. 796–804, 2014.

[7] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 555–573, 2019.

[8] B. L. Due, "Laughing at the robot: Incongruent robot actions as laughables," *Mensch und computer*, 09 2019.

[9] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, 08 2013.

[10] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[11] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, "Evolving trust in robots: Specification through sequential and comparative meta-analyses," *Human Factors*, vol. 63, no. 7, pp. 1196–1229, 2021. [Online]. Available: https://doi.org/10.1177/0018720820922080

[12] J. Lee and K. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, pp. 50–80, 02 2004.

[13] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie, "Transparency in HRI: Trust and decision making in the face of robot errors," *In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, p. 313–317, 2021. [Online]. Available: https://doi.org/10.1145/3434074.3447183

[14] D. J. Brooks, "A human-centric approach to autonomous robot failures," *ProQuest Dissertations and Theses*, p. 229, 2017. [Online]. Available: https://www.proquest.com/dissertations-theses/human-centric-approach-autonomous-robot-failures/docview/1927724896/se-2

[15] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," *Social Robotics*, pp. 42–52, 2017.

[16] C. Bennett, S. Sabanovic, M. Fraune, and K. Shaw, "Context congruency and robotic facial expressions: Do effects on human perceptions vary across culture?" *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, vol. 2014, 08 2014.

[17] M. Otterdijk, E. Barakova, J. Torresen, and M. Neggers, "Preferences of seniors for robots delivering a message with congruent approaching behavior," *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, 07 2021.

[18] K. Malchus, P. Jaecks, O. Damm, P. Stenneken, C. Carles, and B. Wrede, "The role of emotional congruence in human-robot interaction," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 191–192, 03 2013.

[19] C. Esterwood and L. P. Robert, "Do you still trust me? human-robot trust repair strategies," *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 183–188, 2021.

[20] J. B. Lyons, I. aldin Hamdan, and T. Q. Vo, "Explanations and trust: What happens to trust when a robot partner does something unexpected?" *Computers in Human Behavior*, vol. 138, p. 107473, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074756322200293X

[21] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, p. 106561, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563220303101

[22] C. Esterwood and L. P. R. Jr, "Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness," *Computers in Human Behavior*, vol. 142, p. 107658, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563223000092

[23] B. L. Pompe, E. Velner, and K. P. Truong, "The robot that showed remorse: Repairing trust with a genuine apology," *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 260–265, 2022.

[24] S. Jessup, T. Schneider, G. Alarcon, T. Ryan, and A. Capiola, "The measurement of the propensity to trust automation," *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, pp. 476–489, 06 2019.

[25] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies*, vol. 7, pp. 437–454, 11 2006.

[26] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.