The Effect of Voice and Repair Strategy on Trust Formation and Repair in Human-Robot Interaction

The Effect of Voice and Repair Strategy on Trust Formation and Repair in Human-Robot Interaction

MARTA ROMEO, Heriot-Watt University, Edinburgh, United Kingdom of Great Britain and Northern Ireland
ILARIA TORRE, Chalmers University of Technology, Goteborg, Sweden, KTH, Stockholm, Sweden, and University of Gothenburg, Goteborg, Sweden
SÉBASTIEN LE MAGUER, Trinity College Dublin, Dublin, Ireland and University of Helsinki, Helsinki, Finland
ALEXANDER SLEAT, KTH Royal Institute of Technology, Stockholm, Sweden
ANGELO CANGELOSI, The University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland
IOLANDA LEITE, Department of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden

Trust is essential for social interactions, including those between humans and social artificial agents, such as robots. Several factors and combinations thereof can contribute to the formation of trust and, importantly in the case of machines that work with a certain margin of error, to its maintenance and repair after it has been breached. In this article, we present the results of a study aimed at investigating the role of robot voice and chosen repair strategy on trust formation and repair in a collaborative task. People helped a robot navigate through a maze, and the robot made mistakes at pre-defined points during the navigation. Via in-game behaviour and follow-up questionnaires, we could measure people's trust towards the robot. We found that people trusted the robot speaking with a state-of-the-art synthetic voice more than with the default robot voice in the game, even though they indicated the opposite in the questionnaires. Additionally, we found that three repair strategies that people use in human-human interaction (justification of the mistake, promise to be better and denial of the mistake) work also in human-robot interaction.

Marta Romeo and Ilaria Torre contributed equally to this research.

This work was funded and supported by the Manchester–KTH Royal Institute of Technology and Stockholm University Joint Research funding (https://www.manchester.ac.uk/collaborate/global-influence/collaborations/manchester-stockholm/) and partly by the UKRI Node on Trust (EP/V026682/1, https://trust.tas.ac.uk). Aggregated anonymised data and code can be made available upon request to the first two authors.

Authors' Contact Information: Marta Romeo, Heriot-Watt University, Edinburgh, United Kingdom of Great Britain and Northern Ireland; e-mail: m.romeo@hw.ac.uk; Ilaria Torre (corresponding author), Chalmers University of Technology, Goteborg, Sweden, KTH, Stockholm, Sweden, and University of Gothenburg, Goteborg, Sweden; e-mail: ilariat@chalmers.se; Sébastien Le Maguer, Trinity College Dublin, Dublin, Ireland and University of Helsinki, Helsinki, Finland; e-mail: sebastien.lemaguer@helsinki.fi; Alexander Sleat, KTH Royal Institute of Technology, Stockholm, Sweden; e-mail: alexsleat@gmail.com; Angelo Cangelosi, The University of Manchester, Manchester, United Kingdom of Great Britain and Northern Ireland; e-mail: angelo.cangelosi@manchester.ac.uk; Iolanda Leite, Department of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden; e-mail: iolanda@kth.se.

ACM 2573-9522/2025/2-ART33
https://doi.org/10.1145/3711938

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Auditory feedback*; *Empirical studies in interaction design*;

Additional Key Words and Phrases: Trust, Trust repair, Robot voice, Human-Robot Interaction

ACM Reference format:
Marta Romeo, Ilaria Torre, Sébastien Le Maguer, Alexander Sleat, Angelo Cangelosi, and Iolanda Leite. 2025. The Effect of Voice and Repair Strategy on Trust Formation and Repair in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* 14, 2, Article 33 (February 2025), 22 pages.
https://doi.org/10.1145/3711938

1 Introduction

1.1 Trust

Trust is an essential aspect of human-human relationships [19]. Because of its importance, many definitions of trust have been proposed over the years, spanning from evolutionary theories [3], to neurobiological correlates [55], to behavioural economics [28]. In general, *trust* makes someone, the *trustor*, knowingly accept vulnerability to a *trustee*. More formally, trust can be defined as: 'The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party' [45]. From this definition, it emerges that trust is interpersonal and that the trustor has no control over whether their trust has been well-placed or not, so they are exposed to some *uncertainty*. Nevertheless, the trustor has some mental models of how worthy of trust the trustee is [17], and this will influence their decision to trust them or not. As, in the near future, it is conceivable that robots will be used as personal assistants to help people with a wide range of everyday tasks, taking on new roles as social entities, understanding how trust is built and maintained in **Human-Robot Interactions (HRIs)** is of paramount importance. In HRI, one of the most widely used definition of trust is from [41, 50], who define trust as an 'Attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability'. This definition implies both the sense of vulnerability identified in [45] and the acknowledgement that the autonomous agent is a beneficial entity, like in [17].

Moreover, trust is regarded as a multifaceted phenomenon, made of three main components [45]: *Ability*, 'Is the party capable of what they are doing?'; *Integrity*, 'Does the party adhere to a set of acceptable moral principles?'; *Benevolence*, 'Does the party act with good intention without ulterior motives?'. This is particularly important for understanding the main characteristics of a robot that can convey trust. For this reason, it has been revisited to fit the human-agent context in [38], where trust is divided into two different categories: *performance-based trust*, i.e., trust in an agent's ability to complete a task satisfactorily and consistently; and *relation-based trust*, i.e., trust that an agent will comply with social norms. Following this conceptualisation, the **Multi-Dimensional Measure of Trust (MDMT)** scale, consisting of 16 items, has been developed to assess *capacity trust* (composed of reliability and capability factors) and *moral trust* (composed of ethics and sincerity factors) in HRI [79].

1.2 Trust Repair Strategies

Trust between humans and robots can easily be built, due to factors like automation bias [87] or novelty effect [84], but it can be equally easily lost. Typically, once a robot commits a *violation*, trust is lost and needs to be repaired. The severity of the reaction towards the violation, and in turn the efficacy of the *repair strategy* used to re-gain trust, depend on: (i) the type of violation; (ii) the type of task; (iii) the context risk, i.e., how undesirable a failure by the robot would be in

Table 1. Repair Strategies Explored in Human-Human and HRI Literature

Strategy	Scenario	Reference
Apology	Human-human Human-robot	[4, 13, 35, 59, 64]
Denial	Human-human Human-robot	[4, 13, 29, 35, 59, 64]
Trustworthy action	Human-human	[4, 8]
Promising to be better	Human-human Human-robot	[1, 13, 30, 56]
Providing reasons to trust again	Human-robot	[56, 57]
Justify the failure	Human-human Human-robot	[10, 13, 30, 35, 59, 62]
Human support	Human-robot	[8]

that particular context; (iv) the severity of the error/violation [8, 10, 81, 83]. Much effort has been directed into identifying trust-relevant failures leading to violations, and the respective best repair strategies [15, 72]. Generally, the main violations that robots can commit have been classified as *competence-based* and *integrity-based* violations [4, 64], readily mapped to the aforementioned dimensions of trust. Competence-based violations refer to mistakes made by the robot that are task- or hardware-related, like failing to plan or execute an action. Integrity-based violations refer to mistakes made by the robot (intentionally or not) that go against its human partner's principles, like lying or going against established social conventions. The repair strategies adopted by the HRI community to study how robots could re-gain trust are derived from commonly used strategies in human-human interactions and are summarised in Table 1.

In [64], both apology and denial repair strategies are used and compared during a competitive game, with apology ending up being the winning strategy. In fact, in [29] it was shown that having the robot blame anyone (even itself) for a failure caused the users to lose trust in the system. Surprisingly, not many studies have analysed how simply taking proactive and correct actions after a violation influences trust [8]. In [10], justifying the failure, as opposed to ignoring the failure, allowed the robot to mitigate its negative impact in a collaborative game. Similarly, improving the situation awareness of the human partners with respect to the failure and the status of the task being performed by giving support turned out to be an equally winning strategy [8]. The efficacy of the trust repair strategy seems to depend also on the timing of both the violation and the implementation of the strategy. In [12, 57], it was shown that recovery happens slower than the initial trust building and, for this reason, if a failure happens early, trust is heavily negatively impacted. In addition, [56] showed that trust repair strategies implemented immediately after the violation were not effective. Most of the existing studies in the literature focus on one, at most two repair strategies at the time. This is one of the reasons why, although a substantial effort has been now redirected towards understanding repair strategies, to date, what robots should do to recover trust after they commit a violation is still uncertain, as findings are inconclusive [14, 15]. For this reason, in our study we decided to compare three different repair strategies.

1.3 Factors Affecting Trust in Robots

As trust is such a complex construct, many studies exist on investigating the factors that influence trust in the field of HRI. Some of the most comprehensive investigations aiming to identify characteristics that are fundamental for trust are [21, 22]. In particular, they identify three categories

that affect the expression of trust: factors associated with the human, the robot and the context in which the event occurs. These factors can influence each dimension of trust in different ways and a substantial body of works exists that attempts to understand the mechanics of this relationship.

While robot characteristics like anthropomorphism [50] or robot appearance [88] have been investigated as means to convey trust, an important robot-based attribute that has just started to be considered is robot voice [75, 77]. Although there is evidence that voice is a fundamental vehicle for trust [5, 70], it is still a characteristic that gets overlooked by the community. This tendency to overlook voice when designing a robot, or an interaction with a robot, is reflected in the fact that most scholars choose robot voices out of convenience [47], without necessarily thinking of the implications that these voices will have on users' perception. However, this might be problematic because voices can immediately bias people's impressions of their speakers [46]. Voice characteristics such as gender, pitch, accent, speech rate and—in the case of robot voices—degree of human-likeness have all been linked to trust and related phenomena (such as persuasion, attractiveness, agreeableness, sincerity). For example, native accents are often considered more trustworthy than non-native accents [42] and (in the context of the British Isles) regional countryside accents are rated as more trustworthy than regional city accents [7]. High pitch and slow articulation rate can contribute to a voice being perceived as more persuasive [26, 52], although there are floor and ceiling effects: Voices that are too low or high in pitch and speech rate will have the opposite effect [78]. In the context of Human-Machine Interaction, joyful voices are considered more trustworthy than emotionally neutral voices [74]—although this is mediated by the context of the interaction [73]. Trust in artificial voices has mostly been studied in terms of human-likeness [36, 48, 75], and findings are not always in agreement, possibly due to recent advancements in **Text-to-Speech (TTS)** technology. Until recently, this was an issue of balancing intelligibility and flexibility, since it was difficult to generate a clear, understandable artificial voice [39]. However, recent advances in speech synthesis, such as WaveNet, have effectively solved the intelligibility problem, allowing researchers to focus on designing appropriate voices for robots, without having to worry about language comprehension [49]. For this reason, in the current study we chose to compare two different synthetic voices: our robot's default voice (which has been shown to not be considered appropriate for the robotic platform [47]) and another synthetic voice that we generated. Additionally, to the best of our knowledge, no research has investigated the potential intersectional effects of voice and trust repair strategies for HRI. We investigate this in the current article. Specifically, motivated by the fact that voice is a characteristics demonstrated to impact how trustworthy a robot is perceived [61], we investigate whether this perception, conveyed by the differences in the voice, impacts the effectiveness of repair strategies.

Among the three identified categories of [21, 22], context is lately more and more valued as one of the factors that mostly influence trust [9, 61] and this contributes to the difficulties in establishing frameworks and rules surrounding trust in HRI [34]. However, it is clear that, for trust to have an important impact on HRI, the human and the robot need to engage in a collaborative task [22]. Thus, we situated our experimental work in a specific context simulating a human-robot collaborative activity, i.e., a maze solving task.

1.4 Contribution

With our work, we contribute to the empirical literature on trust and trust repair in HRI focusing on robot-related and context-related factors. We compare two different artificial voices and three different repair strategies for competence-based violations and explore their effectiveness on the restoration of trust. To do so, we develop a navigation-based collaborative game (the *maze task*). Navigation tasks have been previously used to investigate trust in HRI [56, 57] and whether unexpected recommendations from a navigation assistant were followed by its human partner

[71]. In particular, mazes offer an intricate and challenging setting and have been used both in psychology [20] and HRI [60] as a way to investigate the development of trust.

In particular, we set off to answer the following **Research Questions (RQs)**:

RQ1: What repair strategies are more appropriate for a collaborative navigation task?
RQ2: Are the preferred repair strategies different for different robot voices?
RQ3: Do the selected repair strategies work in a real-life human-robot collaborative task?
RQ4: Do different combinations of voice and repair strategies influence trust (both behaviourally and perceptually)?
RQ5: Do different combinations of voice and repair strategies influence trust repair after the robot makes a mistake?

To answer these questions, we designed and carried out one online pilot study and one in-person study. The pilot study was used to inform the design of the final in-person experiment, to identify which repair strategy could be used by the robot when it gave a wrong suggestion in a maze-solving problem. Then, we built a physical maze for a social robot (Softbank Robotics's Pepper, shown in Figure 1) to navigate and solve by giving recommendations to participants on which direction to go, and making mistakes along the way. We manipulated the robot's voice and the repair strategy the robot was using to try and re-gain the trust of our participants after it gave a wrong recommendation.

2 Materials and Methods

We conducted an online pilot study to determine which strategy, from the human-human interaction literature, where deemed effective for our task and we then conducted an in-person experiment to study whether this initial result was correct and how it related to a robot characteristic, voice.

2.1 Pilot Study

In the pilot study, we wanted to see what repair strategies people would consider more appropriate for a robot that made a mistake in a maze navigation scenario. To achieve this, we video-recorded the Pepper robot saying six different utterances recommending which direction to take at a junction inside a maze. To increase the believability of the maze context, we edited the videos so that Pepper would appear over a video of a corn maze (see supplementary materials). After watching each video, participants were asked to answer the question: 'Pepper made a mistake. What can Pepper do to regain your trust?'. They could choose from a list of pre-defined trust repair strategies taken from the literature. This list included: apology, denial, promise to be better, failure justification, additional reasons to trust the robot and performing a trustworthy action without explicitly acknowledging the mistake (see Table 2 for the full wording). Note that here they could select more than one repair strategy. Participants watched all six videos, in randomised order. We also asked participants for their gender, age and attitude towards robots.

This study was built on Qualtrics, and participants were recruited using Prolific Academic. We recruited 50 participants, aged 19–70 (median age = 24 years old), of which 28 women, 21 men and 1 non-binary person. The experiment lasted approximately 5 minutes and participants were compensated £1.35. Participants signed a consent form prior to proceeding to the experiment. The experiment was conducted in accordance to ethical guidelines of the hosting institution (KTH Royal Institute of Technology).

2.2 Maze Study

For the main study, we built a maze in a room measuring 4×5 m (Figure 1). We used sheets of blue fabric hanging from the ceiling to create the maze walls. The sheets were hanging about 20 cm

(a) Pepper inside the maze (b) Maze walls

Fig. 1. The maze that was built for the in-person study.

Table 2. List of Trust Repair Strategies and Corresponding Explanations That Participants Could Choose from in the Pilot Study

Repair strategy	Explanation
Implicit improvement	Pepper does not acknowledge in any way the mistake and simply carries on and starts doing the correct thing hereafter.
Apology	Pepper says: 'I'm so sorry I told you to come this way. It's my fault'.
Denial	Pepper says: 'I am sure my sensors readings were correct. It wasn't my fault. I don't know how that happened'.
Promising to be better	Pepper says: 'It won't happen again. I promise to be a better navigator next time'.
Justifying the failure	Pepper says: 'My sensors must have been wrong. It can be due to the nature of the maze and the presence of many walls'.
Giving additional information to trust it again	Pepper says: 'Following my directions will still be faster than trying to guess which direction is the correct one at each crossroads'.

above the floor, so that the robot's safety sensors would not detect the sheets as obstacles, causing the robot to stop. Participants were seated in a room adjacent to the maze room. They were told that they would collaborate with the robot Pepper (which was placed at the entrance of the maze) for this task. The task consisted in finding the 'treasure' (a basket containing golden fabric) hidden inside the maze, and then getting back to the entrance as quickly as possible. Specifically, while the robot physically navigated inside the maze, participants would tell it where to go whenever there was a junction in the path. We decided to separate the robot from the human (and not have them inside the maze at the same time) to add an additional element of vulnerability, which, as mentioned in Section 1, is necessary for trust to emerge [23, 68].

At each junction, the robot told the participants what options were available (e.g., 'Here, we can go left, or straight') and recommended one of the options (e.g., 'I think we should go straight'). Details of how participants communicated with the robot are given below (Section 2.4). To increase the sense of urgency, participants were told that the robot had limited battery autonomy. If participants managed to find the treasure and go back to the entrance before the robot's battery ran out, they were given two supermarket vouchers instead of one (one voucher was worth 50 Swedish crowns). In total, there were nine junctions where a decision from the human participant was required, and a variable number of forced choices depending on which direction participants were guiding the

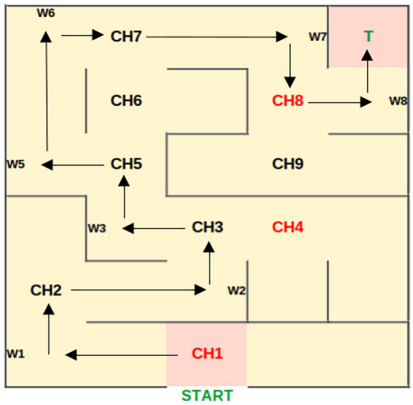

Fig. 2. The maze layout and the optimal path to the treasure 'T'. Junctions where participants had to make a choice are noted as 'CH', the forced choices (the walls) are noted as 'W'. The red 'CH' are the junctions where the robot suggests the wrong path. For example, in 'CH1' the robot suggestion was to go right.

robot (i.e., points where the robot was forced to turn towards a certain direction because of walls). Figure 2 shows the optimal path to the treasure room from the starting point, consisting of six choices and eight forced choices.

The robot was programmed to make up to three mistakes (depending on the navigation plan that the participant decided to follow) at fixed points in the maze. This was done to ensure that participants would hear the robot's trust repair strategy at least once, since participants could navigate in the maze as they wished. In particular, the first mistake was programmed to happen at the very first decision point. A mistake was operationalised as the robot recommending to go towards a direction that would turn out to be a dead-end. If participants followed the robot after this wrong recommendation, the robot would navigate to the location and then say 'Oh no! The road is closed!' or 'Oh no! We went the wrong way!', followed by one of the three repair strategies:

Justify: 'My sensors must have been wrong. It can be due to the nature of the maze and the presence of many walls'.

Promise: 'It won't happen again. I promise to be a better navigator next time'.

Denial: 'I am sure my sensors' readings were correct. It wasn't my fault. I don't know how that happened'.

The experimental procedure was as follows: First, participants were greeted by the experimenters and signed a consent form. Then they completed a questionnaire on their attitudes towards robots (NARS scale). Then they were taken to the entrance of the maze and shown the robot, Pepper (Figure 1). This was done to ensure that all participants had the same mental image of what the robot looked like. However, no interaction with the robot happened at this point as it was shown while being switched off. We explained them that, even if Pepper did not have the complete map of the maze, it was equipped with sensors that help it navigate in unknown environments. In doing so, we wanted participants to understand that Pepper did not have any prior knowledge on the maze but was computing recommendations real-time based on its sensors' readings. This was done to try and create a balanced expectation on the robot capabilities and avoid automation bias effects. Then, demographics data (gender, age, previous experience with robots, spatial orientation skills) were collected. Participants then completed the maze task, which was followed by a questionnaire on their trust in the robot (MDMT questionnaire [79], the current state-of-the-art and validated

questionnaire on human-robot trust). To check the efficacy of our manipulation, we asked two additional questions: 'Do you think Pepper was able to regain your trust after its mistake?' (yes or no question); 'What could it have done instead?' (optional open question). Additionally, we left the possibility to leave any feedback on the experiment. Finally, participants were debriefed and given one or two supermarket vouchers, based on their performance in the maze task. The whole experiment lasted around 30 minutes.

We recruited 60 naive participants using fliers around the university campus and word of mouth. The experiment followed a six conditions between-subject design, each condition being the different combination of voice (Pepper or custom TTS) and repair strategy (Justify–Promise–Denial), resulting in 10 participants per condition. Participants (34 men, 24 women, 2 non-binary) were aged 19–51 (median age = 27); their previous experience with robots was mixed: 7 people reported never having seen a robot before, 20 people regularly watched media (e.g., films, series) with robots, 21 had interacted with a robot before and 12 reported interacting with robots on a regular basis. Among people who had previously interacted with robots, 9 had previously interacted with a Pepper robot. The experiment was conducted in accordance to ethical guidelines of the hosting institution (KTH Royal Institute of Technology).

2.3 Robot Voice Generation

Pepper comes equipped with a default synthetic voice, which sounds 'robotic' and 'child-like' and has been shown to not match Pepper's physical characteristics well [47]. Since the release of this robot and its voice, the speech synthesis technology landscape has been disrupted by the introduction of WaveNet [80] and Tacotron [85], allowing us to synthesise speech with an unprecedented human likeness. For this reason, we decided to compare the influence of Pepper's default voice to another artificial voice that we generated trained using state-of-the-art TTS technologies. Since Pepper's voice is often considered to be feminine, out custom TTS voice was also made to be feminine.

To do so, we relied on the dataset provided for the Blizzard Challenge 2013 [32], comprising of 9,339 utterances extracted from audiobooks, for a total of 19 hours of speech, sampled at 16 kHz, read by a female American Professional narrator.

The TTS system was trained using a combination of three toolkits. First, MaryTTS [66] was used to extract the linguistic features from the text. These features comprise phonetic transcription and elementary prosodic information, consisting of the syllable stress information and the punctuation associated with the pause. Then, FastPitch [37] was used as the acoustic model. While Tacotron [85] is a well-known standard, it provides less control than FastPitch. However, voices generated by FastPitch are judged as equally natural as those produced using Tacotron [40]. Furthermore, with FastPitch we could manipulate duration and fundamental frequency, allowing us to generate sentences that had the same duration and speech rate as the sentences generated with Pepper's default voice. As a final step, WaveNet [80] was used to generate the signal from the mel-spectrogram produced by FastPitch.

The configuration of these toolkits is identical to the one presented in [40], as we used their open source implementations and the provided configuration with limited modifications. The main change relates to the representation used. We used a mel-spectrogram using 80 filters with cutoff frequencies from 50 Hz for the lower bound and 7,600 kHz for the upper bound. We normalised this spectrogram using z-score normalisation as it led to more stable results for both neural vocoders.

As previously mentioned, we used video-recordings of Pepper for the online pilot study. As the videos were not recorded in a sound booth, we post-processed the modern synthetic voice to avoid the acoustic environment influencing the outcome of the experiments. The post-process was achieved using pyroomacoustics [63]. The authors tested different configurations and selected the

Table 3. Acoustic Features of the Two Voices, Pepper (the Pepper Robot's Default Voice) and TTS (Our Custom TTS Voice)

	Pepper	TTS
Mean F0	332.99 Hz	225.67 Hz
SD F0	91.52 Hz	42.93 Hz
Min F0	122.36 Hz	167.47 Hz
Max F0	499.96 Hz	465.89 Hz
Mean HNR	8.61 dB	14.83 dB
Jitter (local)	2.86%	1.33%
Shimmer (local)	15.84%	7.79%

one closest to reference samples. As the duration was imposed during the synthesis stage, the last step consisted of replacing the video audio track with the post-processed speech.

Thus, we ended up with two artificial voices (which we henceforth refer to as 'Pepper's default voice' and 'our custom TTS voice'). Table 3 shows some of the acoustic features that characterise these two voices, taken from the same 13-second utterance spoken by both voices: average F0 and F0 ranges (which correspond to the perceived pitch of a voice [24]) and measures of voice quality (**Harmonics-to-Noise Ratio (HNR)**, and local jitter and shimmer values; these generally indicate how 'clean' and 'sharp' a voice sounds, as opposed to being 'raspy' or 'breathy' [69]). As can be seen from the table, our custom TTS voice has an overall lower pitch than Pepper's voice, while remaining within the average 145–275 Hz female pitch range [27]. The custom TTS voice also has a 'cleaner', less metallic sound (as shown by the voice quality features). Prosodic elements such as duration and speech rate were the same for the two voices; specifically, speech rate was about 3.46 syllables/second—which falls within a normal human speaking rate for conversational English [33].

Qualitatively, Pepper's voice was originally designed to be gender-neutral and child-like (although it is often perceived as being feminine [47]), and it sounds bright and sharp, albeit having a metallic, robotic sound to it. Conversely, the custom TTS voice was made to sound more like an adult woman, with a lower-pitched, husky voice. We wanted a voice that sounded qualitatively different, since most people associate the Pepper robot to female voices.

2.4 Wizard-of-Oz Implementation

During the initial design phase of the maze for the in-person experiment, we envisaged Pepper autonomously navigating within the maze accompanying the participants. However, after building the physical maze we realised that the space left to let both the robot and a person navigate the maze together was too small and could present a safety hazard. We concluded that it would be safer for participants to instruct the robot from outside the maze. This decision allowed us to set up a Wizard-of-Oz experiment in which one of the investigators (the wizard) was teleoperating Pepper to navigate the maze, removing the risks of errors in the navigation. Participants were completing the game from a work-station obtained by enclosing a space in the room adjacent to the maze room, so that they were isolated from the rest of the room and unable to see the wizard and the maze. The work-station setup included a desktop computer, a keyboard, a tablet computer and a pair of noise-cancelling headphones (Figure 3).

The tablet computer was used to initiate a Zoom call with Pepper, so that participants could only see what was immediately in front of the robot during its navigation. Participants had no other information about the maze. The noise-cancelling headphones were connected to the tablet to ensure that participants could clearly hear Pepper's utterances at each junction, while being

Fig. 3. Participant's setup for the in-person study.

shielded from other noises in the room. The desktop computer presented a graphical user interface showing the battery of the robot deprecating over time, a timer, the subtitles to the robot's utterances and the means for participants to make their choices, when they had to. Whenever the wizard navigated Pepper to a junction in the maze, they would activate the TTS engine of the robot, either reproducing a .mp3 audio file for the TTS condition or simply using the in-built function of the robot to reproduce the utterance in the original Pepper voice. At this moment, Pepper would tell participants that it was at a junction and it needed to make a decision on which direction to go, and it would offer its recommendation. Then, participants were asked to confirm their decision using the directional arrows on the keyboard connected to the desktop computer. Participants' computer was connected to the wizard's computer via a peer-to-peer socket communication channel so that, after they selected their choice, the investigator would receive it and navigate the robot accordingly. The wizarded navigation of the robot was also possible thanks to a socket communication channel opened between the wizard computer and Pepper. The wizard simply needed to use the directional arrows on their laptop to generate a message to pass the direction information to the robot, that then translated it into motion.

3 Results

All analyses were conducted in R version 4.3.2, using packages 'tidyverse' (for data cleaning), 'lme4' (for regression analyses), 'ltm' (for internal consistency metrics), 'ggplot2' (for plot generation).

3.1 Pilot Study

In the pilot study, we looked at which repair strategies people considered most appropriate for a robot making a mistake in a maze navigation scenario (via pre-recorded videos; Section 2.1). People's preferences are shown in Figure 4. People's top three preferences of repair strategy were: justification of the mistake, promise to be better, denial. These are therefore the strategies that we implemented in the main in-person study.

3.2 Maze Study

In the main maze study, we looked at a series of behavioural and perceptual measures that gave us an indication of whether people were trusting the robot, and whether this was dependent on the robot's voice and/or repair strategy used.

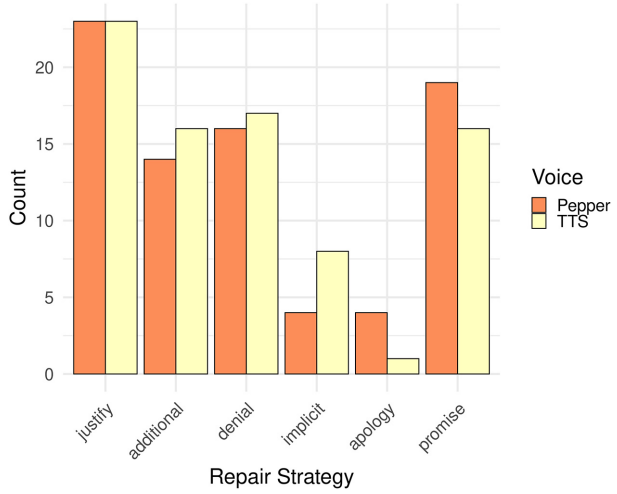

Fig. 4. Repair strategies chosen by participants of the pilot study, in the two voice conditions.

Table 4. Mean Reaction Times (SD in Parentheses), Number of Total Trials Played, and Number of Recommendations Accepted in the Different Voice and Repair Strategy Conditions

Voice	Repair strategy	Reaction time (ms)	# trials	# accepted recommendations
Pepper	Denial	883.42 (1,108.73)	180	145
Pepper	Justify	536.16 (482.02)	167	147
Pepper	Promising	734.65 (880.30)	181	158
TTS	Denial	598.81 (522.21)	157	141
TTS	Justify	523.01 (427.14)	185	161
TTS	Promising	612.02 (783.25)	184	146

Specifically, we looked at the time participants spent taking each decision to accept or not the robot's recommendations. This was calculated as the time between the end of the sentence uttered by the robot and the key press indicating which direction participants wanted the robot to follow. We also looked at the number of times people followed the robot's recommendation. Since the maze contained some forced choices, for this metric we only considered the trials for which participants could actually make a choice. Finally, we looked at participants' answers on the MDMT questionnaire and on the manipulation check questions, which they filled out after the maze task. Summary statistics on the behavioural data can be found in Table 4.

To investigate reaction times, we first removed outliers, i.e., individual trials where participants spent significantly too much or too little time to make a decision. Following established procedures [6], we removed trials where participants spent ±2 SDs away from the mean of all trials reaction times to make a decision. This resulted in the exclusion of 48 trials, for a total of 1,006 analysable trials. Then we built a series of mixed-effects linear models with reaction time as dependent variable, voice and repair strategy as independent variables and participant id as random intercept. The models were built using forward stepwise selection based on the AKAIKE information criterion, and compared to a baseline model using chi-square tests. We found a significant effect of repair strategy ($\chi^2(2) = 6.45, p = 0.039$), with people on average spending less time making a decision in the justify repair strategy than in the denial repair strategy, as can be seen in Figure 5 ($\hat{\beta} = 232.88$, 95% CI $[49.18, 416.59]$, $t = 2.48$).

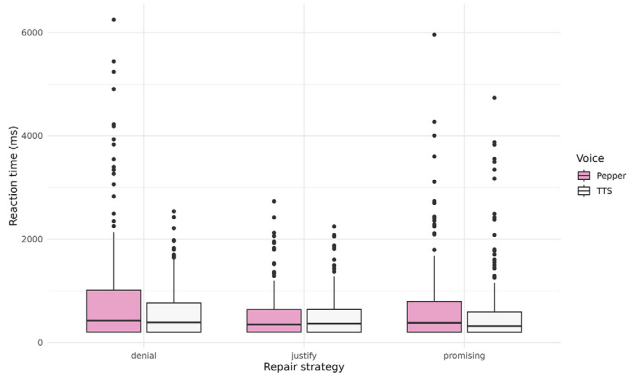

Fig. 5. Average reaction times that participant took to make decisions in the maze, divided by the voice and repair strategy used by the robot.

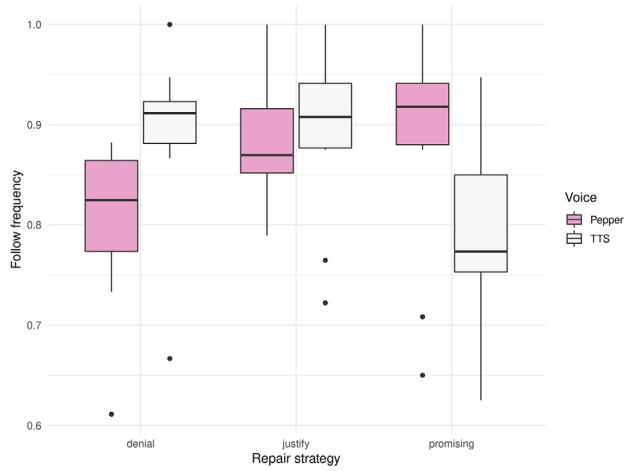

Fig. 6. Frequency of following the robot's recommendations in the maze study, divided by the voice and repair strategy used by the robot.

For the number of times people followed the robot's recommendation, we fitted mixed-effects logistic regression models using participants choice (follow or not) as dependent variable, voice and repair strategy as independent variables and participant id as random intercept. There was a main effect of voice ($\hat{\beta} = 0.76$, 95% CI $[0.07, 1.45]$, $z = 2.15$, p $= 0.031$), with people following the advice of the robot with the TTS voice more. There was no main effect of repair strategy, but there was a significant interaction of voice and repair strategy ($\hat{\beta} = -1.37$, 95% CI $[-2.31, -0.43]$, $z = -2.86$, p $= 0.004$), with people following the robot with TTS voice and promising strategy the smallest number of times, as shown in Figure 6.

Due to the nature of our experimental design, people could have encountered a different number of robot errors—and thus heard a different number of repair strategy utterances—based on which route they chose to follow. Of our 60 participants, the vast majority (n = 35) encountered two robot errors, n = 15 encountered 1, n = 8 encountered 3, n = 1 encountered 4 and n = 1 did not encounter any error. As an exploratory analysis, we fitted another mixed-effects logistic regression model on participants' choice to follow or not the robot, with the additional explanatory variable of number

Fig. 7. Mean ratings from the MDMT questionnaire, divided by the voice and repair strategy used by the robot.

of encountered errors. This was done mainly to see whether the addition of this variable removed the effect of voice from our main model described above, and not to see whether the number of errors could have influenced participants' decisions (since this was not a controlled manipulated variable). Adding this co-variate did not change our main results, and the same main effect of voice and interaction effect of voice and repair strategy were found.

To examine whether people's trust and behaviour towards the robot changed after there was a trust breach—i.e., after the robot made a mistake—we split the game data into before and after hearing the second repair utterance from the robot, and conducted the same analysis (regressing the number of times people followed the robot) with the split (before/after the second mistake) as an additional predictor. We did not consider the reaction times before and after the split because there would most likely be a confound with people's progressively getting faster as they learn how the game works. We split the data after the second repair utterance, and not the first, since the first one happened at the very first choice in the maze. Only participants who encountered at least two robot mistakes (n = 44) were included in this part of the analysis.

We found that people followed the robot on average the same number of times before and after witnessing the second mistake: $\hat{\beta} = -0.16$, 95% CI $[-0.60, 0.28]$, $z = -0.72$, p = 0.472.

Finally, for the MDMT questionnaire, we first calculated internal consistency scores for each of the original four subscales of 'reliable', 'capable', 'sincere', 'ethical' [79]. We found good to excellent consistency for each of them (Cronbach's Alpha = 0.77, 0.93, 0.79, 0.80, respectively). Then, we conducted ANOVA tests with the two questionnaire scores (capacity and moral trust) as dependent variables, and robot voice and repair strategy as independent variables (Figure 7). For capacity trust, there was no main effect of voice or repair strategy, but there was a significant interaction ($F(2, 54) = 6.02$, $MSE = 1.45$, p = 0.004, $\hat{\eta}^2_G = .182$), with people giving higher ratings to the Pepper's default voice and justify repair strategy combination. Similarly, there were no main effects for moral trust, but the same interaction was found ($F(2, 52) = 3.31$, $MSE = 2.20$, p = 0.044, $\hat{\eta}^2_G = 0.113$). One of the peculiarities of the MDMT questionnaire is that it gives the possibility to respondents to indicate that a certain item 'Does not fit' the robot in question. This is done to avoid forced and meaningless ratings. If 'Does not fit' is checked, the item becomes a missing value. Hence, our participants had the chance to select 'Does not fit' whenever they did not consider an item relevant. However, this option was rarely selected (only two participants selected 'Does not fit' enough for their Moral Trust scale to result in missing and were thus omitted from the analysis), and data points were collected consistently for both scales.

To the manipulation check question 'Do you think Pepper was able to regain your trust after its mistake?' out of our 60 participants 53 (88%) answered positively, while the remaining 7 answered negatively. Only 18 participants (30%) specified what the robot could have done to regain their trust, and 30 (50%) left additional comments. To analyse the qualitative data collected through the open questions, two of the authors independently coded the answers and confronted them to find general trends. Although not many generalisable concepts were found in the answers, what was clear from their breakdown was that we could identify two main takeaway messages: (1) participants understood that Pepper made a mistake; (2) some participants had personal biases or personality traits that were the main cause for not complying with the robot's suggestions. In fact, we find that only three participants commented with variations of 'I do not think it made a mistake', while the rest understood that the robot was sometimes at fault. Thirteen participants suggested that the robot could have given more information, four participants clearly referred to the fact that Pepper should have 'Taken Responsibility' and 'Apologise', four participants wished that Pepper was faster in both navigating and delivery the repair strategy (i.e., 'It should move on faster. Just a short sorry and then continue'). Finally three participants directly stated that their personal attitudes (i.e., 'being too stubborn') were what influenced their choices during the Maze Task.

4 Discussion

With the studies presented here, we looked at whether a robot's voice influences trust formation and repair in a collaborative task, and whether a series of repair strategies commonly used in human-human interaction also work when the trust is broken by a robot.

Stemming from definitions of trust commonly used in the HRI literature [22, 23, 68], we created a scenario where people and robot had to cooperate to solve a problem, whereby both had access only to partial information. We also introduced a vulnerability component, which is fundamental to trust, in that participants' monetary reward at the end of the study was dependent on how well they solved the task.

In the post-task questionnaire, we found that participants rated Pepper's default voice as more trustworthy than our custom TTS voice. However, this is not reflected in people's behaviour in the maze task, where they actually trusted the robot with our custom TTS voice more—as indicated by the average number of times that people decided to follow the robot's advice. This apparent discrepancy actually confirms research in Psychology showing that people's explicit judgments do not often correlate with their behaviour [18]; in other words, people don't do what they say they would do. A recent experiment in HRI also found that people's perception of a robot's competence-based trust did not correlate with any of the collected behavioural measures [67]. This has important implications for HRI, because it suggests that different robot characteristics—in this case, voice—might be more or less appropriate for different contexts and interaction modalities. Based on our current result, we hypothesise that Pepper's default voice—which is rather high-pitched and expressive, almost girlish or child-like—might be preferable in general contexts, where no trust action is required from the human user. On the other hand, our custom TTS voice—which is lower pitched and sounds like a rather serious adult woman—might be preferable in contexts where the human user has to accept a certain risk and needs to perform an action, such as deciding whether to accept the robot's recommendation or not. Our results confirm other recent evidence that different artificial agent features and behaviour might be more effective in different contexts: For example, [73] showed that smiling, up-beat avatars acting as navigation assistants in a hypothetical lunar crash scenario were not trusted as much as avatars showing a neutral, 'serious' expression. Similarly, robots that actively nudged people to take action in a simulated evacuation scenario were more effective than robots that simply awaited human instructions [25]. This also opens new avenues for future work. Apart from keeping investigating which voice is most appropriate for which context

(see, e.g., [76]), what would the best course of action be for a robot that needs to interact with humans in different contexts? Should it change its voice to better fit the current interaction context, or would this change in a robot's persona result in an Uncanny Valley?

Given that context is crucial [61], with our online pilot study we wanted to find the repair strategy/ies that was considered most suitable for our scenario (RQ1). We were also interested in investigating whether the different characteristics of the robot were influencing the decision on the repair strategy in any way (RQ2). No effect of voice was found (RQ2), but there was a general consensus that justifying the failure, promising to do better and even denying the mistake were acceptable repair strategies in our chosen cooperative context (RQ1). The results on the justification and promise strategies are consistent with previous findings [10, 65]. However, the positive results of the denial strategy were unexpected; from previous studies, it had emerged that having the robot blame anything else but itself leads to a decrease in trust [29], and that denial was the least successful strategy in repairing trust in a human-robot teaming scenario. Additionally, while previous studies showed that apology is an acceptable strategy in some contexts [56, 64], it did not fare well in our joint navigation scenario. This might be because the nature of the trust-requiring tasks in these studies was different: For example, [64] compared denial and apology in a competitive setting (whereas our task was cooperative in nature), and in [56] the high-risk evacuation scenario resulted in the understanding that the strategy was successful at repairing trust if used when the robot asks the human to trust it again, but not when used immediately after the mistake. Once again, we find that trust-related variables are highly context-dependent, and this extends to trust-repair strategies as well. In the maze study, we found some evidence that the justify strategy might have been more immediately convincing to participants, since they made up their mind more quickly after hearing this strategy, as compared to the denial strategy (RQ3). Since the repair strategies did not significantly differ in terms of probability of following the robot's recommendation however, it is possible that participants needed less time to think about what to do next, when the robot provided an explanation for its mistake, even though the following action was not affected. On the other hand, people hearing the denial strategy might have needed some time to think of a possible reason behind the mistake, even though their final decision was the same. Thus, the explicit choices that people made in the pilot study were generally reflected in the behavioural results in the maze task, in the sense that all repair strategies worked equally well. There was only one exception, in that the combination of our own TTS voice and the promise repair strategy was particularly unsuccessful (RQ5). This supports previous research on the topic, which argues that agent features (such as voice and body shape [47], facial expressions and linguistic content [2], facial and vocal expressivity [73]) should not be examined in isolation, because their various combinations might be perceived differently. The promise repair strategy had also already been found to be less effective in HRI in the case of repeated robot errors [51]. This might be due to the underlying moral implications of breaking a promise previously made.

Similarly, the repair strategies that were adopted by the robot did not have an overall effect on the perceived trustworthiness of the robot in the post-game questionnaire. However, Pepper's default voice combined with the justify repair strategy gave rise to higher ratings of trust in the MDMT questionnaire (RQ4). From this, we can infer that the perceived trustworthiness of the robot could be a factor influencing the effectiveness of the repair strategy but not the other way around. In fact, what we found with our maze experiment was that no repair strategy was more successful than the others in boosting the trustworthy perception of the robot and guiding participants towards trusting the robot after a violation (as the number of times they followed the robot in the experiment did not depend on the repair strategy).

Finally, we found that people's behaviour before and after witnessing the robot's mistakes was essentially unchanged, as indicated by the number of times people decided to accept the robot's

recommendations (RQ5). This means that the trust breach was successfully addressed by the repair utterances spoken by the robot—as confirmed by participants in their answers to the post-task questionnaire. Thus, we do not find one trust repair strategy to work better than others overall, but rather all three investigated repair strategies worked in this cooperative context.

However, even if the mistake and trust repair efforts of the robot were acknowledged, it is still possible that our participants did not consider the robot's mistakes to be crucially trust-breaking. While similar navigation errors were used in previous studies [57, 58], and while we tried to increase participants' vulnerability with a potential monetary penalty, our study scenario is low-stakes. Vulnerability is a fundamental factor for trust to emerge [68] and it is deeply linked to the notion of risk [54]. For this reason, since the risks for the participants of our study were not high, it is possible that the trust-breaking moment did not impact them strongly. This might also be the reason why using the Denial strategy was not detrimental to the perception of the trustworthiness of the robot, as this strategy has been shown to work best when the exact causes of the trust violations are unknown [43]. In the future, to be able to generalise our results, a more high-stake scenario should be designed and tested.

In human-human interaction, denial has been shown to work as a trust repair strategy for integrity-based trust violations, but not competence-based ones [16, 31]. As an example, in [31], participants were shown videos of potential job candidates who had committed trust violations at their previous job (compiling tax declarations incorrectly). The violation was manipulated to be either competence-based (filling out the wrong form) or integrity-based (knowingly providing the wrong information). The candidate responded with a trust-repairing utterance (either apologising, or denying the accusation). Participants then rated the candidate for perceived trustworthiness. Results showed that apology was a better repair strategy for the competence-based violation while denial was a better strategy for the integrity-based violation. In our study, while we did not explicitly tell participants what kind of trust violation the robot committed, the nature of the task pointed to a competence-based violation scenario, since it was both in the human and robot's interest to successfully complete the task. Speculating that participants understood this, we can suggest that some strategies that worked to repair trust in HRI (such as denial in competence-based violation scenarios) might not work in human-human interaction [16, 31], or the other way around.

Overall, we found that robot voice influenced trust formation, and henceforth its repair, while repair strategy did not strongly affect the human-robot trust relationship in this cooperative setting. This finding contributes to the ongoing research on trustworthy autonomous systems. We confirm that the literature on trust in HRI is inconclusive not because it does not provide guidelines to design social robots that could convey trust; on the contrary, it is hard to find a general definition of a trustworthy robot, as the context in which the robot is to be deployed plays a fundamental role in how the robot should behave, appear and sound to convey and repair trust. Lastly, we report findings that suggest that new research efforts should be directed toward understanding what strategies robots could use to regain trust after they commit a violation. While it is necessary to get inspired by the human world, it is becoming more and more clear that a simple transposition of what we do in our everyday interactions does not completely translate to robots when it comes to trust [14]. For this reason we should start investigating *ad-hoc* methods for HRI, as previously suggested by [4].

4.1 Limitations and Future Work

This work presents some limitations that should be acknowledged. One is the low-risk scenario that we devised for our behavioural trust measurement, which we have addressed above.

Another limitation is the relatively low sample size of our experiment. While the total number of in-person participants was 60 (a considerable number overall), we ended up having only 10

participants for each of the experimental conditions, due to the high number of conditions we tested and the between-subject experimental design. This was due to logistical and institutional limitations. One user session took roughly 1 hour, and the experimenter needed an additional half an hour between sessions to deal with potential robot-related issues (charging, switching off to prevent over-heating, etc.). Furthermore, due to the nature of the hosting institution (which is a technical university and does not offer, e.g., courses in Psychology), recruitment was cumbersome, since there is no existing set-up where students can sign up to participate in experiments in exchange for credits. This means that our population was a bit more varied, but at the cost of availability and time spent. We do not wish to endorse publishing underpowered studies, but we acknowledge that conducting user studies with robots in person requires significant efforts in terms of time and resources, meaning that convenience sampling needs to be adopted, in line with existing norms in the HRI literature [44, 53, 82, 86].

Also, due to the nature of the maze task, people encountered a different number of decision points, and witnessed a different number of robot mistakes, depending on the route they chose to take. This means that their individual experience was not fully controlled and might have led to a few artefacts in the analyses. However, this was necessary to ensure the ecological validity of the task. We strove to give our participants an experience as close to reality as possible, taking inspiration from collaborative scenarios in which humans and robot need to find a solution together and where humans have real choices over the decisions to be made. For this reason, the path they could take was not pre-planned, and we believe that this facilitated the observed behavioural differences. However, as described in Section 3.2, we took this into account in our analyses (by adding the number of encountered errors as an explanatory variable, and by performing an analysis only on participants who encountered two or more errors) and we did not find any influence of these variables on the overall results. However, future studies should seek to confirm and expand our results with a higher sample size and a more controlled experimental setup.

In addition, due to the physical limitations of the robot (which walks at a slower-than-average human walking speed) and the room where we built the maze (which is rather small), we could not have the robot and human collaboratively solve the maze together in the same room. This would be worth exploring in the future, as the proximity to a physical robot, as well as different robot embodiments, may play a role in the overall participants' experience and robot perception [11].

Finally, our work mainly focused on competence-based violations, and the repair strategy was employed immediately after the robot made a mistake. One additional variable that would be worthy of investigation in future works is the timing of the repair strategy and how that influences compliance with the robot recommendations (in line with the work of [65]).

References

[1] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Investigating the effects of (empty) promises on human-automation interaction and trust repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20)*. ACM, 6–14. DOI: https://doi.org/10.1145/3406499.3415064

[2] Markus Appel, Birgit Lugrin, Mayla Kühle, and Corinna Heindl. 2021. The emotional robotic storyteller: On the influence of affect congruency on narrative transportation, robot perception, and persuasion. *Computers in Human Behavior* 120 (2021), 106749.

[3] Annette Baier. 1986. Trust and antitrust. *Ethics* 96, 2 (1986), 231–260. DOI: https://doi.org/10.1086/292745

[4] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. 2018. Toward an understanding of trust repair in human-robot interaction. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (2018), 1–30. DOI: https://doi.org/10.1145/3181671

[5] Pascal Belin, Bibi Boehme, and Phil McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS One* 12, 10 (2017), e0185651.

[6] Alexander Berger and Markus Kiefer. 2021. Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology* 12 (2021), 675558.

[7] Hywel Bishop, Nikolas Coupland, and Peter Garrett. 2005. Conceptual accent evaluation: Thirty years of accent prejudice in the UK. *Acta Linguistica Hafniensia* 37, 1 (2005), 131–154.

[8] Daniel J. Brooks, Momotaz Begum, and Holly A. Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 487–492. DOI: https://doi.org/10.1109/ROMAN.2016.7745162

[9] David Cameron, Jonathan M. Aitken, Emily C. Collins, Luke Boorman, Adriel Chua, Samuel Fernando, Owen McAree, Owen Martinez Hernandez, and James Law. 2015. Framing factors: The importance of context and the individual in understanding trust in human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2015*. Retrieved from https://eprints.whiterose.ac.uk/91238/

[10] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, 507–513.

[11] Nathaniel Steele Dennler, Stefanos Nikolaidis, and Maja Matari'c. 2024. Singing the body electric: The impact of robot embodiment on user expectations. arXiv:2401.06977. Retrieved from https://api.semanticscholar.org/CorpusID:266999021

[12] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 251–258. DOI: https://doi.org/10.1109/HRI.2013.6483596

[13] Connor Esterwood and Lionel P. Robert. 2021. Do you still trust me? Human-robot trust repair strategies. In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 183–188. DOI: https://doi.org/10.1109/RO-MAN50785.2021.9515365

[14] Connor Esterwood and Lionel P. Robert. 2022. A literature review of trust repair in HRI. In *Proceedings of the* 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, 1641–1646. DOI: https://doi.org/10.1109/RO-MAN53752.2022.9900667

[15] Connor Esterwood and Lionel P. Robert Jr. 2023. Three strikes and you are out! The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142 (2023), 107658.

[16] Matteo Fuoli, Joost van de Weijer, and Carita Paradis. 2017. Denial outperforms apology in repairing organizational trust despite strong evidence of guilt. *Public Relations Review* 43, 4 (2017), 645–660.

[17] Diego Gambetta. 1988. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*. Diego Gambetta (Ed.), Basil Blackwell, 213–237.

[18] Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102, 1 (1995), 4.

[19] Herbert P. Grice. 1975. Logic and conversation. In *Speech Acts*. Peter Cole and Jerry L. Morgan (Eds.), Brill, 41–58.

[20] Joanna Hale, Madeleine E. M. Payne, Kathryn M. Taylor, Davide Paoletti, and Antonia F. De C. Hamilton. 2018. The virtual maze: A behavioural tool for measuring trust. *Quarterly Journal of Experimental Psychology* 71, 4 (2018), 989–1008. DOI: https://doi.org/10.1080/17470218.2017.1307865

[21] Peter A. Hancock, Deborah Billings, Kristin Schaefer, Jessie Chen, Ewart de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53 (Oct. 2011), 517–27. DOI: https://doi.org/10.1177/0018720811417254

[22] Peter A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, John C. Brill, and James L. Szalma. 2021. Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors* 63, 7 (2021), 1196–1229. DOI: https://doi.org/10.1177/0018720820922080

[23] Glenda Hannibal, Astrid Weiss, and Vicky Charisi. 2021. "The robot may not notice my discomfort" – Examining the experience of vulnerability for trust in human-robot interaction. In *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 704–711. DOI: https://doi.org/10.1109/RO-MAN50785.2021.9515513

[24] Daniel J. Hirst and Céline de Looze. 2021. Measuring speech. fundamental frequency and pitch. In *Cambridge Handbook of Phonetics*. Rachael-Anne Knight and Jane Setter (Eds.), Cambridge University Press, 336–361. DOI: https://doi.org/10.1017/9781108644198

[25] Yuhan Hu, Jin Ryu, David Gundana, Kirstin H. Petersen, Hadas Kress-Gazit, and Guy Hoffman. 2023. Nudging or waiting? Automatically synthesized robot strategies for evacuating noncompliant users in an emergency situation. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 603–611.

[26] Margarete Imhof. 2010. Listening to voices and judging people. *The International Journal of Listening* 24, 1 (2010), 19–33.

[27] J. Jackson and R. D. A. Taylor. 2019. Vocal pitch and intonation characteristics of those who are gender non-binary. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS '19),* 2685–2689. International Phonetic Association, London, UK. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_2734.pdf

[28] Gareth R. Jones and Jennifer M. George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of Management Review* 23, 3 (1998), 531–546.

[29] Poornima Kaniarasu and Aaron M. Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 850–855. DOI: https://doi.org/10.1109/ROMAN.2014.6926359

[30] Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. 2023. "What if it is wrong": Effects of power dynamics and trust repair strategy on trust and compliance in HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23).* ACM, 271–280. DOI: https://doi.org/10.1145/3568162.3576964

[31] Peter H. Kim, Donald L. Ferrin, Cecily D. Cooper, and Kurt T. Dirks. 2004. Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology* 89, 1 (2004), 104.

[32] Simon King and Vasilis Karaiskos. 2013. The blizzard challenge 2013. In *The Blizzard Challenge Workshop*. Retrieved from http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf

[33] Xaver Koch and Esther Janse. 2016. Speech rate effects on the processing of conversational speech across the adult life span. *The Journal of the Acoustical Society of America* 139, 4 (04 2016), 1618–1636. DOI: https://doi.org/10.1121/1.4944032

[34] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.

[35] Johannes Maria Kraus, Julia Merger, Felix Gröner, and Jessica Pätz. 2023. 'Sorry' says the robot: The tendency to anthropomorphize and technology affinity affect trust in repair strategies after error. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23).* ACM, 436–441. DOI: https://doi.org/10.1145/3568294.3580122

[36] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Frontiers in Neurorobotics* 14 (2020), 593732.

[37] Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 6588–6592. DOI: https://doi.org/10.1109/ICASSP39728.2021.9413889

[38] Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2021. The interplay between emotional intelligence, trust, and gender in human–robot interaction: A vignette-based study. *International Journal of Social Robotics* 13, 2 (04 2021), 297–309. DOI: https://doi.org/10.1007/s12369-020-00624-1

[39] Sébastien Le Maguer and Benjamin R. Cowan. 2021. Synthesizing a human-like voice is the easy way. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, 1–3.

[40] Sébastien Le Maguer, Simon King, and Naomi Harte. 2022. Back to the future: Extending the blizzard challenge 2013. In *Proceedings of Interspeech 2022*, 2378–2382. DOI: https://doi.org/10.21437/Interspeech.2022-10633

[41] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. DOI: https://doi.org/10.1518/hfes.46.1.50/_30392

[42] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology* 46, 6 (2010), 1093–1096.

[43] Brinsfield C. and Lewicki R. J. 2017. Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior* 4 (2017), 287–313.

[44] Joe Louca, Kerstin Eder, John Vrublevskis, and Antonia Tzemanaki. 2024. Impact of haptic feedback in high latency teleoperation for space applications. *Journal of Human-Robot Interaction* 13, 2, Article 16 (June 2024), 21 pages. DOI: https://doi.org/10.1145/3651993

[45] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734. DOI: https://doi.org/10.2307/258792

[46] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One* 9, 3 (2014), e90779.

[47] Conor McGinn and Ilaria Torre. 2019. Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* ACM, 211–221.

[48] Wade J. Mitchell, Kevin A. Szerszen Sr, Amy Shirong Lu, Paul W. Schermerhorn, Matthias Scheutz, and Karl F. MacDorman. 2011. A mismatch in the human realism of face and voice produces an Uncanny Valley. *i-Perception* 2, 1 (2011), 10–12.

[49] Roger K. Moore. 2017. Appropriate voices for artefacts: Some key insights. In *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*.

[50] Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. ACM, 33–42. DOI: https://doi.org/10.1145/3319502.3374839

[51] Birthe Nesset, Marta Romeo, Gnanathusharan Rajendran, and Helen Hastie. 2023. Robot broken promise? Repair strategies for mitigating loss of trust for repeated failures. In *Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1389–1395. DOI: https://doi.org/10.1109/RO-MAN57019.2023.10309558

[52] Oliver Niebuhr, Alexander Brem, Eszter Novák-Tót, and Jana Voße. 2016. Charisma in business speeches: A contrastive acoustic-prosodic analysis of Steve Jobs and Mark Zuckerberg. In *Proceedings of the 8th Speech Prosody Conference*. Speech Prosody Special Interest Group, 79.

[53] Denis Peña and Fumihide Tanaka. 2020. Human perception of social robot's emotional states via facial and thermal expressions. *Journal of Human-Robot Interaction* 9, 4, Article 26 (May 2020), 19 pages. DOI: https://doi.org/10.1145/3388469

[54] Brianna J. Tomlinson, Rachel E. Stuck, and Bruce N. Walker. 2022. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science* 23, 4 (2022), 500–516. DOI: https://doi.org/10.1080/1463922X.2021.1975170

[55] René Riedl and Andrija Javor. 2012. The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics* 5, 2 (2012), 63.

[56] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2015. Timing is key for robot trust repair. In *Proceedings of the International Conference on Social Robotics (ICSR)*. Springer International Publishing, 574–583.

[57] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2017. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436. DOI: https://doi.org/10.1109/THMS.2017.2648849

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

[59] Kantwon Rogers, Reiden John Allen Webber, and Ayanna Howard. 2023. Lying about lying: Examining trust repair strategies after robot deception in a high-stakes HRI scenario. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. ACM, 706–710. DOI: https://doi.org/10.1145/3568294.3580178

[60] Marta Romeo, Peter E. McKenna, David A. Robb, Gnanathusharan Rajendran, Birthe Nesset, Angelo Cangelosi, and Helen Hastie. 2022. Exploring theory of mind for human-robot collaboration. In *Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 461–468. DOI: https://doi.org/10.1109/RO-MAN53752.2022.9900550

[61] Marta Romeo, Ilaria Torre, Sébastien Le Maguer, Angelo Cangelosi, and Iolanda Leite. 2023. Putting robots in context: Challenging the influence of voice and empathic behaviour on trust. In *Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2045–2050.

[62] Andres Rosero. 2023. Using justifications to mitigate loss in human trust when robots perform norm - Violating and deceptive behaviors. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. ACM, 766–768. DOI: https://doi.org/10.1145/3568294.3579979

[63] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A Python package for audio room simulation and array processing algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 351–355.

[64] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I don't believe you": Investigating the effects of robot trust violation and repair. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 57–65. DOI: https://doi.org/10.1109/HRI.2019.8673169

[65] Sonja Stange and Stefan Kopp. 2021. Explaining before or after acting? How the timing of self-explanations affects user perception of robot behavior. In *Proceedings of the 13th International Conference on Social Robotics (ICSR)*. Springer, 142–153.

[66] Ingmar Steiner and Sébastien Le Maguer. 2018. Creating new language and voice components for the updated MaryTTS text-to-speech synthesis platform. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 3171–3175.

[67] Rebecca Stower, Karen Tatarian, Damien Rudaz, Marine Chamoux, Mohamed Chetouani, and Arvid Kappas. 2022. Does what users say match what they do? Comparing self-reported attitudes and behaviours towards a social robot. In *Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1429–1434. DOI: https://doi.org/10.1109/RO-MAN53752.2022.9900782

[68] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178–186.

[69] João Teixeira and André Gonçalves. 2014. Accuracy of jitter and shimmer measurements. *Procedia Technology* 16 (Dec. 2014), 1190–1199. DOI: https://doi.org/10.1016/j.protcy.2014.10.134

[70] Arnaud Tognetti, Valerie Durand, Melissa Barkat-Defradas, and Astrid Hopfensitz. 2020. Does he sound cooperative? Acoustic correlates of cooperativeness. *British Journal of Psychology* 111, 4 (2020), 823–839.

[71] Hiroyuki Tokushige, Takuji Narumi, Sayaka Ono, Yoshitaka Fuwamoto, Tomohiro Tanikawa, and Michitaka Hirose. 2017. Trust lengthens decision time on unexpected recommendations in human-agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction (HAI '17)*. ACM, 245–252. DOI: https://doi.org/10.1145/3125739.3125751

[72] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. ACM, 3–12. DOI: https://doi.org/10.1145/3319502.3374793

[73] Ilaria Torre, Emma Carrigan, Katarina Domijan, Rachel McDonnell, and Naomi Harte. 2021. The effect of audio-visual smiles on social influence in a cooperative human–agent interaction task. *ACM Transactions on Computer-Human Interaction* 28, 6 (2021), 1–38.

[74] Ilaria Torre, Jeremy Goslin, and Laurence White. 2020. If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior* 105 (2020), 106215.

[75] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *Proceedings of APAScience '18: Technology, Mind, and Society (TechMindSociety '18)*, 1–6. DOI: https://doi.org/10.1145/3183654.3183691

[76] Ilaria Torre, Adrian Benigno Latupeirissa, and Conor McGinn. 2020. How context shapes the appropriateness of a robot's voice. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 215–222. DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223449

[77] Ilaria Torre and Laurence White. 2021. Trust in vocal human–robot interaction: Implications for robot voice design. In *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*. Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, and John J. Ohala (Eds.), Springer, Singapore, 299–316. DOI: https://doi.org/10.1007/978-981-15-6627-1_16

[78] Ilaria Torre, Laurence White, Jeremy Goslin, and Sarah Knight. 2023. The irrepressible influence of vocal stereotypes on trust. *Quarterly Journal of Experimental Psychology* 77 (2023), 17470218231211549.

[79] Daniel Ullman and Bertram F. Malle. 2019. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 618–619.

[80] Aäron Van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. In *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 125.

[81] Sanne van Waveren, Elizabeth J. Carter, and Iolanda Leite. 2019. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19)*. ACM, 151–158. DOI: https://doi.org/10.1145/3308532.3329475

[82] Alessia Vignolo, Henry Powell, Francesco Rea, Alessandra Sciutti, Luke Mcellin, and John Michael. 2021. A humanoid robot's effortful adaptation boosts partners' commitment to an interactive teaching task. *Journal of Human-Robot Interaction* 11, 1, Article 9 (Oct. 2021), 17 pages. DOI: https://doi.org/10.1145/3481586

[83] Jie Wang, Wuji Lin, Xu Fang, and Lei Mo. 2020. The influence of emotional visual context on the judgment of face trustworthiness. *Psychology Research and Behavior Management* 13 (2020), 963.

[84] Yawei Wang, Qi Kang, Shoujiang Zhou, Yuanyuan Dong, and Junqi Liu. 2022. The impact of service robots in retail: Exploring the effect of novelty priming on consumer behavior. *Journal of Retailing and Consumer Services* 68 (2022), 103002. DOI: https://doi.org/10.1016/j.jretconser.2022.103002

[85] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: A fully end-to-end text-to-speech synthesis model. arXiv:1703.10135. Retrieved from http://arxiv.org/abs/1703.10135

[86] Steve Whittaker, Yvonne Rogers, Elena Petrovskaya, and Hongbin Zhuang. 2021. Designing personas for expressive robots: Personality in the new breed of moving, speaking, and colorful social home robots. *Journal of Human-Robot Interaction* 10, 1, Article 8 (Feb. 2021), 25 pages. DOI: https://doi.org/10.1145/3424153

[87] Julia L. Wright, Jessie Y. C. Chen, Michael J. Barnes, and Peter A. Hancock. 2016. The effect of agent reasoning transparency on automation bias: An analysis of response performance. In *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality*. Springer, 465–477.

[88] Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F. Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics* 7, 1 (2016), 000010151520160005. DOI: https://doi.org/doi:10.1515/pjbr-2016-0005

Received 16 May 2023; revised 04 November 2024; accepted 29 November 2024