GRIFFITH UNIVERSITY

An attempt to reduce negative attitude towards robots and upgrade trust through explanations

Author

Javaid, M, Estivill-Castro, V, Hexel, R

Published

2019

Conference Title

Australasian Conference on Robotics and Automation 2019, ACRA 2019 Proceedings

Version

Version of Record (VoR)

Rights statement

© 2019 Australian Robotics and Automation Association. The attached file is reproduced here in accordance with the copyright policy of the publisher. Please refer to the conference's website for access to the definitive, published version.

Downloaded from

http://hdl.handle.net/10072/401211

Link to published version

https://ssl.linklings.net/conferences/acra/acra2019_proceedings/views/by_auth.html

Griffith Research Online

https://research-repository.griffith.edu.au

An Attempt to Reduce Negative Attitude Towards Robots and Upgrade Trust Through Explanations

Misbah Javaid, Vladimir Estivill-Castro and Rene Hexel
School of Information and Communication Technology
Griffith University, 4111, QLD, Australia
misbah.javaid@griffithuni.edu.au

Abstract

When humans interact with robots in daily life, each human has a different attitude towards robots that may directly affect human-robot trustworthy relationships. By attitude, we mean any mental disposition matured through experience. A negative attitude is the psychological factors that prevent a human from interacting with robots in daily life and also creates a hurdle for a human to build trust in a robot. In this paper, we hypothesise that explanations from a robot that contain *"Decision-Transparency"* and *"Error-Justification and Correction"* policy can help in reducing humans' negative attitude towards robots and facilitate smooth interaction. Explanations, specifically communicated *in human-understandable terms* can create a significant difference. To analyse the profound impact of explanations from a robot, we conducted an Experimental Study with 34 human participants by performing a decision-making task in collaboration with a real robot. Objective assessment, i.e. facial expressions and eye contact with the robot signalled a decrease in the negative attitude of human participants towards the robot. We also found that human participants trusted and conformed more with the robot's decisions (communicated in terms of explanations), as compared to their own decisions. Meanwhile, subjective measures (Negative Attitude toward Robots Scale (NARS), Human-Robot Trust Scale (HRT) questionnaires) also reported that, after having interaction with the robot, humans' trust in the robot increased and negative attitude significantly reduced. Our findings suggest new implications for the establishment of smooth human-robot trustworthy relationships.

1 Introduction

The primary focus in the world of robotics is currently placed on the design and capabilities of the technology itself. The robotic industry has moved towards robots which are both social and functional [Shaw-Garlock and Glenda, 2009] and conversely, the human element has been overlooked and ignored. However, if the human factor of the human-robot pair is neglected in the design and implementation of the technology component, different challenges will inevitably emerge for the creation and validation of broad-spectrum of successful interactions between humans and robots. Through a vigilant survey of the literature, we identified a growing number of investigations, and empirical explorations highlight different aspects that influence interactions involving humans and robots [Yagoda *et al.*, 2012]. One of the leading issues is the question of trust of a human interacting with a robot.

For decades, trust has been investigated in many ways (i.e., interpersonal trust, trust in automation). However, still, there is much space to study trust that humans attribute to robots during Human-Robot Interaction. Trust decides the overall acceptance of a system. For this reason, there is a considerable number of studies seeking a better understanding of the human-robot trust dynamics [Hancock *et al.*, 2011]. Another factor to consider is the negative attitude of humans towards robots, which is a psychological factor and needs special consideration because it prevents individuals from interacting with robots and consequently creates hurdles for humans to build trust on robots. Nevertheless, several studies have validated a significant set of benchmarks for social acceptance of technology, i.e., performance expectancy, humans' attitude towards technology [Heerink *et al.*, 2012]. In case of robots, to develop efficient human-robot interaction also demands (1) the acceptance of robots in our society, (2) humans' positive attitude towards robots and (3) trust by the humans. Therefore, humans' trust and attitudes towards robots can be a more eloquent indicator for daily and sustained

human-robot interactions [Gaudiello *et al.*, 2016]. Attitude is basically a firm and enduring tendency to behave or react in a specific way towards other humans or objects and is caused by different perspectives like cultural background, and personal experience [Chaplin and James Patrick, 1968].

This definition provides us with the motivation that the attitude of humans towards robots can be altered by focusing on some factors like personal experience. We hypothesise that, based on human-specific needs and objectives, if the robot provides explanations in *human-understandable terms*" that can help in reducing the influence of negative attitude on humans' behaviours toward robots, and upgrade humans' sense of trust in a robot. In general, explanations are given to impart, modify or to clarify knowledge [Nothdurft *et al.*, 2014] and serve to make something clear and understandable. With explanations capability, the behaviour of a robot becomes more understandable for humans that will contribute to reducing humans' negative attitude toward the robot and will establish a human-robot trustworthy relationship. The main motivation behind this study is that robots have become increasingly central to our society, and humans' interaction with them is evolving from *master-slave* to *peer-to-peer*.

Therefore, it seems timely and important to understand how to promote their interactions with humans. Humans with a highly negative attitude toward robots mentally tend to trust robots less and avoid communication with robots. We should aim to investigate methods that are help reduce the negative attitude of humans for short-term and long-term social interactions with robots. Therefore, we suggest a more human-like approach by augmenting a robot with the explanations.

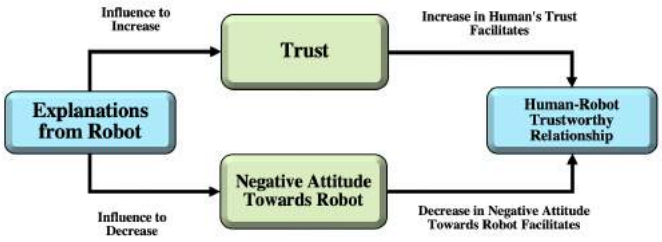


Figure 1: Proposed Research

To investigate the profound impact of robot explanations, we carried out an *Experimental Study* with 34 human participants and a real robot. We created different image flashcards characterised by different *Queensland (QLD) Traffic Regulations*, *Traffic Signs* and *Road Problem Solving* scenarios - more like (*Hazard Perception*) scenarios. Human participants interact with the robot individually to complete a decision-making task with the robot.

The decision-making task is tricky, and *time-sensitive*, in which the human and the robot have to solve different *Road Problem-Solving* scenarios, by deciding in an uncertain environment, within the given time (150 seconds for each scenario). We also inform the human participants that this task requires collaboration with the robot, and the final decision does not depend on the human's decision solely. We explained to every human participant that first, the human participant must select an option, next participants listen to the robot's explanations and are allowed to change their answer or to leave it as it was. For the proposed method, we set the focus of our inquiry through humans' *conformation* [Gaudiello *et al.*, 2016] and acceptance to robot's answers, as a new objective measure of the human-robot trust relationship.

The reasoning engine on board of a robot could make mistakes or could be prone to errors (because of lack of information or heuristic reasoning). Depending on the nature of the task, errors can have an adverse effect on humans' trust and even can contribute to increasing the negative attitudes of humans. From a performance standpoint, if robots can notice their errors and recover themselves automatically, robots will be considered more reasonable and reliable by humans. Learning from mistakes will ultimately increase humans' willingness to interact with the robot in the future. In real life, we are tempted to trust humans and generate positive vibes for them if they could explain to us why they do what they do. This retraction is how often explanations shall generate empathy. Witnessing humans realisation of their mistakes and explain how they correct their failures to us is a factor to build trust. We suggest humans have the same expectation from social robots. Therefore, we manipulated the robot to make a mistake by generating *Wrong Explanations*[1] and immediately correct itself with a sophisticated justification for the error. This study sets out to understand the influence of the *Error-Justification and Correction* policy by which a robot can recover from a failure in order to mitigate its possible adverse social effects. We investigate whether, during the *time-sensitive task*, when the robot justifies the failure to a human team-mate and corrects itself, will retraction mitigate the negative effects of the erroneous situation and reduce humans' negative attitude. We note that in some context, recognising a mistake is perceived as naturally human [Crigger, 2004]. The follow-up question is whether humans will re-build their trust on the robot. By addressing these questions, this work contributes to the design metrics of future robots, respectively.

The literature regarding erroneous behaviours of so-

[1]A *Wrong Explanation* means the robot contradicts the scenario with certain reasons and produces *Wrong Explanations* intentionally.

cial agents and robots is still very recent [Shiomi, Nakagawa and Hagita, 2013]. Some researchers began to inquire about strategies to not only cope with the erroneous behaviours but also to mitigate their adverse effects [Brooks *et al.*, 2016]. However, previously recovery strategies were tested through online surveys, without creating a direct interaction between a human and a faulty robot. One of the disadvantages of conducting an online survey for the perception of a robot is that participants act only as observers. As such, they are not directly affected by the robot's presence and interaction. This study avoids this issue as the human participants have *peer-to-peer* interaction with a robot in the same environment. Humans may be directly affected by the robot's proficient explanations.

To assess the attitude of humans towards our robot, we use the most common attitude scale NARS (Negative Attitudes towards Robots Scale) [Nomura *et al.*, 2006] because it focuses mainly on communicative robots. Within the scope of our study, our robot is also equipped with communication ability in terms of explanations. The NARS is composed of three different types of subscales of negative attitudes. *NARS-S1* deals with the *negative attitude towards situations of interaction with robot*, *NARS-S2* describes *negative attitude towards social influence of robots*, *NARS-S3 negative attitude towards emotions in interaction with robots* [Gaudiello *et al.*, 2016]. The primary goal of NARS is to predict human interaction based on humans' negative attitudes towards robots [Nomura *et al.*, 2006]. *NARS-S1* and *NARS-S2* have a particular focus on the negative attitude towards interaction and social influence of robots, and they are especially relevant with relation to trust in human-robot interaction socially. In this sense, a negative attitude can influence humans' mistrust in the possibility that a robot might fit interactions in the social structures. Therefore, we expect that the more human will show negative attitudes towards social interaction of robots, the less they trust the robot's decisions concerning social interactions. Furthermore, *NARS-S1* and *NARS-S2* are also relevant to our methodological choice of employing explanations and affect trust. It is reasonable to think that the more a human will feel anxious with the idea of being influenced by a robot (in terms of getting and accepting explanations from a robot), the less a human will accept and change its decisions according to robot's decisions.

Human participants' trust is not directly observable by using a 14-items subscale of the Human-Robot Trust Questionnaire [Schaefer and Kristin, 2013] because that questionnaire focuses specifically on the robot's functional capabilities, before interaction and after interaction as *pre-study*, *post-study* questionnaires. In this paper, we explore the effect of explanations in terms of *"Decision-Transparency"* and *Error-Justification and Correction*[2]. Our focus is reducing humans' negative attitude towards robots and upgrading humans' sense of trust in a robot. To our knowledge, this issue has not yet been explored.

We divided this paper into different sections. Section 2 surveys the literature regarding trust and explanations in the context of HRI. Section 3 provides a brief description of our experiment material. Section 4 discusses the *Case Study* in detail along with the hypothesis, design and procedure of experiment and measurement of dependent variables along with human recruitment and participation. Section 5 presents obtained results in detail in the light of our proposed hypothesis. Finally Section 6 presents a discussion and Section 7 describes conclusion keeping in view the implications of this work for the HRI community.

2 Related Work

For decades, trust has been investigated in many ways (i.e., interpersonal trust and trust in automation) but still, there is much space to study trust that humans attribute to robots in HRI. There have been a growing number of investigations and empirical explorations regarding different factors that affect human-robot interaction [Yagoda *et al.*, 2012]. Hancock and colleagues [Hancock *et al.*, 2011] reported 29 empirical studies. They developed a triadic model of trust as a foundation to provide a greater understanding of different factors that facilitate the development of humans' trust in robots. As a focus of interest, the model generally features robot-related factors, environmental-related factors and human-related factors [Schaefer and Kristin, 2013]. Each of these factors plays a significant role in the development of humans' trust in robots [Hancock *et al.*, 2011]. However, robot-related characteristics [Hancock *et al.*, 2011] primarily associated with robot capabilities include etiquette in a robot (remained aware of its mistakes) influence humans' trust most dramatically. Most of the previous investigations on the effect of explanations used logic-based Artificial Intelligent (AI) systems and rule-based systems (especially within the context of expert systems [Swartout *et al.*, 1993]). Early work in the area of expert systems suggested that automatic generation of explanations improved acceptability of these systems. Also, systems which provided some explanations after some failure received more tolerant behaviour by humans. Intelligent systems (neural networks, case-based reasoning systems, and heuristic expert systems) have also explored using an explanation facility [Darlington and Keith, 2013]. However, for human-robot interaction, there has been little empirical evaluation of the impact

[2]As a strategy of recovery from a mistake.

of explanations on humans' trust. Dzindolet [Dzindolet *et al.*, 2003] investigated manually crafted explanations and showed hand-crafted explanations to be promising in providing transparency and improved trust. However, hand-crafted explanations were static and manually created, fell short of transferring the complexity of robotic decision-making to human teammates. Nothdurft [Nothdurft *et al.*, 2014] research on explanations and trust kept as the primary focus the trustworthy relationship between human and computer. Robot capabilities include robot performance (i.e., performance) are the most critical factors, to date, in an occurrence of trust development [Hancock *et al.*, 2011]. Robot's functional capabilities are crucial to facilitate appropriate human-robot interaction. The importance of functional capabilities on trust development is also one of the most well-researched areas of human-robot trust [Schaefer and Kristin, 2013]. Finally, concerning our work, the act of providing explanations from a robot is also associated with the robot's functional capabilities because providing explanations is a *Robot-Related Factor*.

3 Experiment Material

We created 105 flashcards containing different images regarding *Traffic Regulations*, *Traffic Signs* and *Road Problem-Solving* scenarios[3] on it. We created three possible types of flashcards (35 of each type): Figure 2 shows an example of each type of flashcards.

Figure 2: Images on the flashcards - (a) **TYPE - 1** (*Traffic Regulation*), (b) **TYPE - 2** (*Traffic Sign*), (c) **TYPE - 3** (*Road Problem Solving - Hazard Perception Scenario*)

3.1 TYPE - 1 (Flashcards with Traffic Regulations)

Type 1 flashcards contains only *Traffic Regulations*, written in text format (See Figure 1 (a)) The primary goal with the *Type-1* flashcard is, we want a human to observe the reading ability[4] and capability of the robot to produce correct and relevant explanations according to the image on the flashcard. Expected explanations from the robot regarding Figure 2 (a).

"A motorcycle passenger, that sits behind the rider is called a pillion passenger. Listen human! A pillion passenger should only ride on the motorcycle when the motorcycle has a suitable pillion seat along with suitable passenger footrests. Remember this as well human, all of the passengers must also wear an approved motorcycle helmet, securely fastened, that complies with Australian standards AS 1698".

3.2 TYPE - 2 (Flashcards with Traffic Signs)

Type 2 flashcard contains only *Traffic Symbol* on it (See Figure 1 (b)). The robot sees the flashcard, recognises the traffic sign and produces relevant explanations to the human. The primary goal with the *Type-2* flashcard is we want a human to analyse that the robot is not only capable of reading, but it also has a correct assessment of the *Traffic Sign* and consequently producing relevant explanations according to its assessment. Expected explanations from the robot regarding Figure 2 (b).

"This is Australian Disability Parking Scheme permit. In Queensland, holders of this permit are entitled to park in any space provided for a person, with a disability, in an on-street, or off-street parking location, for example, shopping centres, and hospitals. Disability parking permits are issued and maintained by the Department of Transport and Main Roads Queensland Australia."

3.3 TYPE - 3 (Flashcards with Road Problem Solving Scenarios)

Type 3 flashcard contains *Road problem-solving scenario* more like a *Hazard Perception Scenario* on it (see Figure 1 (c)). We created *Road Traffic problem-solving* scenarios, which are more towards solving the problem by making decisions in uncertain road situations and communicating those decisions to humans in terms of relevant explanations. The primary goal associated with the *Type - 3* flashcards is, we want a human to assess that the robot not only correctly assesses the *Traffic Sign* and explains well a *Traffic Regulation* but it can do something meaningful in a complicated situation. The human should not only be convinced that the robot knows the *Traffic Rules*, but should also be convinced that the robot has some sound judgements to break a *Traffic Rule* in order to keep a road situation safe. Expected explanations from the robot regarding Figure 2 (c).

"According to the given scenario, the white vehicle, which is vehicle A, can not cross the double lines, except

[3] All the images and explanations are created according to the *QLD Department of Transport and Main Roads (TMR)*. See `www.tmr.qld.gov.au` for more details.

[4] Correctly reading without making any mistake.

to safely pass a cyclist. However, it is a new law, as if a human is driving at 60 kilometres per hour, and a cyclist is passing by the human. It is his responsibility to maintain a distance of 1 meter, according to his current speed, or 1.5 meters, and remember human, if the human increases his car speed, there is a penalty of 287 dollars for drivers, plus 60 dollars, victims of crime levy, and also 2 demerit points; so be careful. If in doubt, it is better to hold back, until there is a safe spot, to overtake them. Hope you understand human !"

4 Case Study

To investigate the profound effect of explanations in reducing humans' negative attitude and establishing a human-robot trustworthy relationship, we pose the following hypothesis :

4.1 Hypothesis

- **Hypothesis 1**: Explanations from a robot in human-understandable terms serve to reduce negative attitude of a human towards the robot and upgrade human's sense of trust in a robot.
- **Hypothesis 2:** In an uncertain environment, explanations from a robot, that disseminate *Error-Justification and Correction* (after the faulty behaviour), help to reduce the adverse effect of the erroneous situation and support *Hypothesis 1.*
- **Hypothesis 3**: Human participants accept and conform more with the robot's decisions (communicated in terms of explanations), as compared to their own decisions, when confronted with uncertainty in the environment, during a decision-making task.

4.2 Design of Experiment

We used a between-subjects design for our *Experimental Study*, in which human participants have to interact with a robot within two possible conditions :

- **Condition 1** - *Control Condition*: The robot makes no mistake and provides *Correct Explanations.*
- **Condition 2** - *Error-Justification and Correction*: The robot makes a mistake (intentionally, but that is not apparent to the human participant) and provides *Wrong Explanations*, but immediately corrects itself with a sophisticated justification.

During the decision-making task, the robot does not answer which option is correct. Instead, the human participants must use their common sense to verify the correct option by listening carefully to the robot's explanations. In *Condition 2*, to justify it's failure, the robot generates different words along with gestures for example: "*I am sorry for wrong assessment*" scratching its head to show that it is recalling the Correct Explanation, "*oh wait human partner, let me have a look again,*" and "*Sorry I disagree with you human, my belief is ...*" (see Figure 3 (a), (b)). In this way, we also assessed the impact of *acceptance of mistake* towards human participants.

Figure 3: After a faulty behaviour, the robot corrected itself. (a) The robot is scratching his head and recalls the *Correct Explanation.* (b) The human participants are listening to the robot carefully by maintaining an eye contact. (c) The human participant is looking at the robot with strange facial expressions.

Humans' Conformation to Robot as an Innovative Measure of Human-Robot Trust:
In the HRI research field, many mechanisms to measure humans' trust on robots have been developed, but these are mostly based on self-report (questionnaires). These measures reflect a human's specific mental posture concealed in an apparent and clear opinion; therefore, it is difficult to analyse those spontaneous opinions; mostly based upon a human's inner belief and are limited in their capacity to analyse further on which robot knowledge, a human has built its trust. One complementary approach in this perspective is *Media Equation Theory* [Nass *et al.*, 2000], which illustrates that, when humans engage in collaborative tasks with computers, they tend to accept computers as social entities unconsciously. Therefore, they trust answers provided by the computer and conform their answers according to it. We adapted the famous *Media Equation Theory* paradigm for our study. During the decision-making task (which is characterized by uncertainty), we measured human participants' conformation to the robot's answers; generated in terms of explanations, to specific questions as an innovative measure of human's trust on the robot. In particular, we examined (1) humans' trust in the robot competency by assessing its correct situation awareness and change their answers after getting explanations from the robot, (2) or reject the robot answers and stick with their answer(s). The robot did not give answer directly by telling which option is correct[5] but provides explana-

[5]The human participants have to use their common sense to select or verify the correct option by listening carefully to the *Robot Explanations.*

tions to make it's decisions transparent according to the given road situation and also to justify the cause if there is a failure.

4.3 Procedure of Experiment

In the literature, human-robot trust has been measured either implicitly by contrasting objective measures or explicit by using subjective measures. Objective measures are retrieved from behavioural data (i.e., response time) unconsciously produced by individuals while subjective measures deal with self-reports and questionnaires retrieved from collected verbal data consciously produced by the individuals [Gaudiello *et al.*, 2016]. The former approach is limitedly developed in HRI, and the latter is widely used. This research adopted the approach of combining survey(s) with an experiment to evaluate humans' trust towards the robot. We performed our experiment in three-stages.

Stage - 1 of the Experiment

During *Stage 1*, we evaluated human participants' initial level of trust towards the robot, by filling the HRT questionnaire[6].

Stage - 2 of the Experiment

Stage 2 has several steps. First, the human participant selects freely six flashcards (three of each type, i.e. TYPE 1, TYPE 2 and TYPE 3) and shows them to the robot sequentially and listen to the explanations. Following this, the human participants perform the decision-making task with the robot, as per the following sub-steps. (1) Human participants focus on the three flashcards (containing *Road Problem-Solving Scenario*) from the pile of flashcards. (2) Meanwhile, a *Monitor Screen* displays the scenario on the screen, the human participant can review the screen and selects an answer. (3) After Step 2, the human participants can listen to the robot explanations again. Following this, they can change their answer or not. The answer will be saved. Every human participant is given 150 seconds for each scenario.

Figure 4: Human participants are showing flashcard to the robot.

[6]Pre-interaction questionnaires are essential because these are filled before interaction with the robot and provide us with the expected mental model of human participants towards the robot.

Figure 5: The human participants are selecting their answer according to the given *Traffic Scenario*.

Stage - 3 of the Experiment

Trust is a dynamic attitude that changes over time [Schaefer and Kristin, 2013]. On the completion of the experiment with the robot, as a possible clarification of change in the level of trust towards the robot, human participants' filled a second HRT questionnaire[7]. Change in the level of trust will help us in examining the profound influence of the explanations from the robot.

4.4 Dependent Variables Measure

The dependent variables fall into two categories:

1. Negative attitude towards robots will be assessed through the use of most common attitude scale NARS (Negative Attitudes towards Robots Scale) [Nomura *et al.*, 2006].
2. Human participants' trust will be analysed by using Human-Robot Trust Questionnaire [Schaefer and Kristin, 2013].

4.5 Recruitment and Participation

This study was conducted in an Australian university, and there was a total of 34 human participants, (16 females and 18 males) with age ranging from 18 to 35 years old ($M = 18.2 \pm 4.59$). Since this was an individual activity, we kept a balance of human participants in each condition (17 in *Condition 1* and 17 in *Condition 2*). We recruited participants through general advertisement, using posters on university notice board and also with direct contact with students, and by word of mouth. We gave an incentive of AUD $10 gift card as a token of appreciation for their active participation.

5 Experiment Results

In this section, we present the results of the subjective and objective assessments of the effect of *Robot Explanations* on human participants' level of trust, set in the context of the human-robot collaborative scenario, using the statistical software package *SPSS for Mac*. Before conducting any analysis, we performed a reliability

[7]Post-interaction questionnaires provide us with the actual mental model of human participants towards the robot. The comparison of the expected mental model with that of actual mental model reports the difference between these two.

analysis (Cronbach's α) to assess the internal reliability of the HRT [Schaefer and Kristin, 2013] and NARS [Nomura *et al.*, 2006] Questionnaires. HRT had $\alpha > 0.723$ and NARS had $\alpha > 0.714$ respectively.

5.1 Impact On Trust

After reliability analysis, we performed a normality analysis using *Shapiro-Wilk Test*, to examine whether the depend variable representing trust follows a normal distribution for the groups based upon *Condition 1* and *Condition 2*. The test reported a normal distribution for both groups.

Condition - 1 (No-Error)

We performed parametric *paired sample t-test* to analyse the overall effect of the robot explanations. We compared human participants level of trust towards the robot after interaction with the robot, controlling for the trust levels reported before interaction. Results showed a significant difference (t (16) = -7.512, p<0.001), suggesting that the dependent samples t-test is appropriate in this case. Figure 6 (a), shows a glimpse of effect of *Robot Explanations*, that reflects significant higher trust levels towards the robot after interaction ($M = 96.41 \pm 4.63$), when compared to their trust level with that of before interaction ($M = 62.76 \pm 7.69$).

Condition - 2 (Error-Justification and Correction)

With the help of parametric *paired sample t-test*, we analysed the overall effect of the robot's faulty behaviour on a human's level of trust. We examined human participants' level of trust towards the robot when it produced an error but corrected itself immediately with that of before interaction with the robot. The results showed a significant difference (t (16) = -22.50, p<0.001), suggesting the suitability of dependent samples t-test. Figure 6 (b), illustrates the results with bar graph, that reflects significant higher trust levels towards the robot after interaction ($M = 83.06 \pm 8.52$), when compared to their trust level with that of before interaction ($M = 56.63 \pm 6.19$).

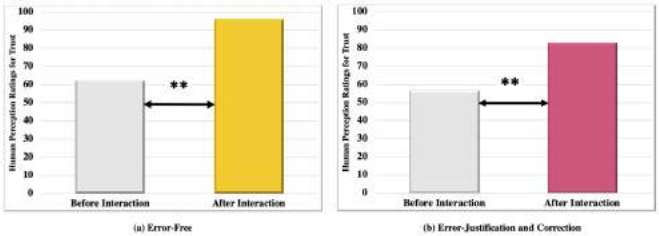


Figure 6: Difference in the trust level of human participants before and after interaction with the robot (**Correlation is significant at $p < 0.01$)

5.2 Impact On Negative Attitude Towards Robot

After a reliability analysis, we performed a normality analysis using *Shapiro-Wilk Test*, to examine whether the dependent variable representing negative attitude follows a normal distribution for the groups based upon *Condition 1* and *Condition 2*. The test reported that NARS (sub-scale S1 and sub-scale S2) does not follow a normal distribution for both groups. Therefore, we performed a (non-parametric) *Wilcoxon signed-rank test* to analyse[8] any difference between "*Negative Attitude*" before interaction and after interaction with the robot. *NARS-S1* mainly deals with the negative attitude towards situations of interaction with the robot. *NARS-S2* mainly concerns the negative feelings of humans that arise negative attitude towards the social influence of robots.

Condition - 1 (No-Error) :

The *Wilcoxon* test revealed significant difference (Z = -3.524, p<0.01) for *NARS-S1* and (Z = -3.413, p=0.01) for *NARS-S2*, after having the interaction with the robot. Figure 7 (a) and (b) reflect a decrease in the *NARS-S1* and *NARS-S2* ratings when the behaviour of the robot was *Error-Free*. This means the negative feelings of human participants (with this perception that the robot can impact or control them, and they will be confused or depend on the robot remarkably for making decisions in an uncertain environment) significantly reduced after interacting with the robot.

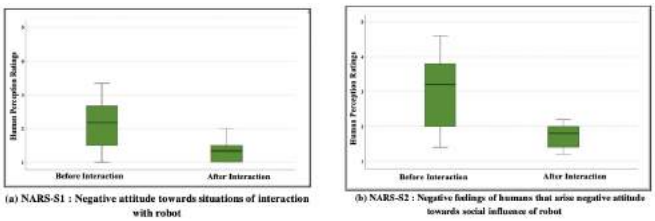


Figure 7: Significant decrease in the ratings of NARS S1 Subscale and NARS S2 Subscale; when the behaviour of the robot was *Error-Free*.

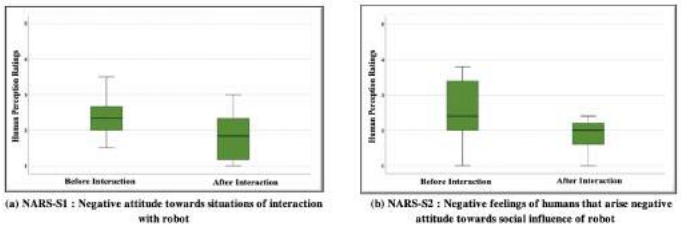
Condition - 2 (Error-Justification and Correction)

The *Wilcoxon* test revealed significant difference (Z = -2.109, p<0.05) for *NARS-S1* and (Z = -2.438, p=0.01) for *NARS-S2*, after interaction with the robot.

Figure 8 (a) and (b) show a decrease in the ratings of *NARS-S1* and *NARS-S2*, even when the behaviour of

[8]We computed the scores for NARS-S1 and NARS S2 by the method recommended by authors [Nomura *et al.*, 2006]. NARS-S1 score ranges from minimum 6 to maximum 30, and NARS-S2 score ranges from minimum 5 to maximum 25.

the robot was erroneous. This is because of the *Error-Justification and Correction* policy, which supports our idea that if the robot not only perceives its mistake but also accept them with a decent self-justification, this also helps in reducing the negative feelings of humans towards robots.



(a) NARS-S1 : Negative attitude towards situations of interaction with robot

(b) NARS-S2 : Negative feelings of humans that arise negative attitude towards social influence of robot

Figure 8: Even with the faulty behaviour of the robot, still there is a decrease in the ratings of NARS S1 Subscale and NARS S2 Subscale; because the robot corrected itself immediately by giving a sophisticated justification for the error.

5.3 Human Participants Conformation to the Robot

The human participants selected choices for the scenarios were recorded in a *CSV* file for further analysis. These answers helped us to create quantitative measures of human participants' trust in the robot's functional capability,[9] that will ultimately help in reducing negative attitudes of the human participants towards the robot. Our method to calculate the conformation score was to divide the number of times a human participant changed his answer according to the robot's answer by the total number of times where the robot's answer mismatched with the human participant's first answer. Therefore, we build a reasonable score for the analysis ranging between 0 (no conformation) to 1 (full conformation). We considered a score greater than or equal to 0.5 as a human participant's trust and acceptability of the robot.

	Mean	Standard Deviation
Explanations With No-Error	0.62	0.314
Explanations with Error-Justification and Correction	0.6	0.42

Figure 9: Conformation score for the group of human participants, $N = 17$ in each group under Condition 1 and Condition 2.

Interestingly, human participants were willing to accept and conformed more to the robot's answers as compared to their answers. We examine whether a group of human participants under *Condition 1* conformed more with the robot or the group of human participants under *Condition 2*. Descriptive analysis was performed to analyse the normal distribution of the conformation score, which revealed that the conformation score is not normally distributed. Therefore, we performed a (non-parametric) *Wilcoxon Test* for paired samples, which indicated no significant difference between the two groups (Z = -0.224, p>0.05).

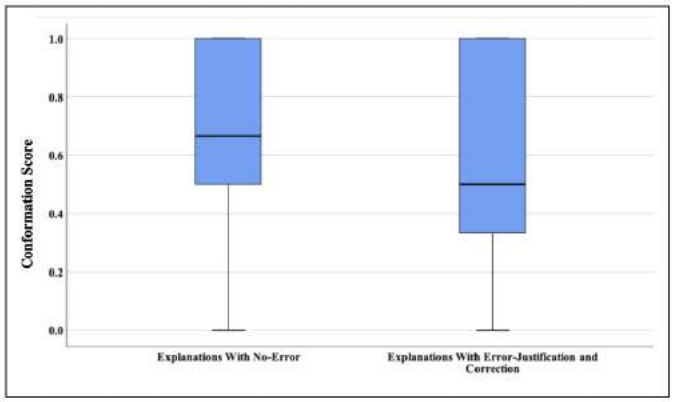


Figure 10: Human participants' conformation with the robot's decisions

Even when human participants were sure about their selected choices, still they verified them by asking for help from the robot, as shown in Figure 11. Maybe the human participants rely on the robot more to take decisions because they may observe and interpret that the robot has some criteria or logical demonstration to apply its knowledge of *Traffic Rules* and does not apply them blindly. Even, if the robot considers breaking a *Traffic Rule*, this is also based on some evaluation of the danger of the situation.

Figure 11: Decision-making task - After selecting the option(s), the human participants are verifying their selected option by asking advice from the robot.

6 Discussion

Overall, the results provide strong support to our **Hypothesis 1**, by showing a decrease in the ratings of the negative attitude of human participants and an increase in the level of trust towards the robot. To test our **Hypothesis 2**, we manipulated the robot to contradict with

[9] If human participants changes their answer so many times after getting explanations from the robot, then we can say that the humans trust the robot.

the image or situation in the *Problem-Solving* scenario with specific reasons and produced *Wrong Explanations*. However, by adopting the *Error-Justification and Correction* policy, the negative impact of the erroneous situation did not produce serious consequences. Our primary purpose was not only to check the impact of faulty behaviour but also to analyse if the explained withdraw sustains trust. We programmed the robot to display a behaviour where it appears to discover its error, accepts its mistake and explains it (by saying sentences such as "*I am sorry*," and "*Wait a bit human, I overlooked the situation, let me analyse it again.*" We investigated whether it make any difference on the perception of the human participants. We found this has an impact and it was reflected during the decision-making task, where the human participants accepted and conformed more with the robot's decisions as compared to their answers. This supported our **Hypothesis 3**. Besides, ratings of the *NARS-S1* subscale and *NARS-S2* were also negatively correlated with the conformation score in both conditions. This negative correlation also reflects that the human participants conformed more with the robot's decisions.

However, human participants did not conform to the robot's answers easily. First of all, explanations from the robot helped the human participants to scrutinize the quality of the information provided by the robot. Therefore, the human participants had a fair understanding that the robot has not only a reasonable knowledge of *Road Rules* but also has excellent *Road Problem-solving* skills, as well as it is prepared to respond accordingly. Secondly, many human participants showed the same flashcard to the robot more than once to confirm whether the robot gives the same explanations again or not. Because, we created three sets of explanations for each flashcard, and the explanations given the first time did not trigger again. Thirdly, we also noticed that human participants, after looking at the *Monitor Screen*, also scrutinize the flashcard to examine the image and to inspect whether the *Robot Explanations* are aligned with the image on the flashcard as shown in Figure 12. We

Figure 12: The robot is giving explanations and the human participants are looking into flashcards to scrutinize whether the robot explanations are aligned with the image on the flashcard.

wanted this to happen because the human participants should notice that the robot does not explain randomly, and it has some criteria to apply the *Traffic Rules*, by not applying them blindly. We also observed human participants' utterance, especially when the robot reflected on a mistake but corrected itself. Participants showed admiration: *wao*, *genius*, and *intelligent*. As another signal of willingness to interact that also influences trust, they also maintained eye contact with the robot. Collectively, all these factors showed that the human participants understood and recognised the robot as an expert, which was reflected when human participants' withdraw their answers and conformed more to the robot's answers during the decision-making task. This constitutes an appropriate measure to straightforwardly registering the human participants' trust in the robot. Overall, the negative attitude of human participants towards the robot significantly reduced, and trust of the human participants significantly increased after having the interaction with the robot, in both conditions, i.e., *No-Error* condition and *Error-Justification and Correction* condition. However, the human participants in the group where the robot did not have erroneous behaviour showed higher trust levels and very low ratings of negative attitude towards the robot.

7 Conclusion

We should aim to investigate methods that help reduce humans' negative attitude towards robots and establish human-robot trustworthy relationship. Several studies have also validated a significant set of indicators of social acceptance of technology such as performance expectancy, humans' attitude towards technology [Heerink *et al.*, 2012]. These indicators are useful to estimate humans' willingness to accept robots. Especially during social interactions in which a human relies on a robot to make decisions or exchange information. We are also interested in identifying a set of factors which are likely to correlate with our work. For example, individuals' fear of being influenced by robot [Nomura *et al.*, 2006], which causes different behaviour of individuals towards the robot. In this perspective, we implemented an HRI scenario in which a robot was equipped to provide explanations in *human-understandable terms* by recognising and explaining different *Traffic Rules* and *Traffic Signs*. Moreover, the robot has the ability to solve different *Road-Problem Solving* scenarios by making decisions in uncertain road situations. During the design process, we make sure that the scenario should introduce some moments of distrust so that we could quantify the differential impact of *Error-Justification and Correction* policy on a human's level of trust. Overall, we analysed that the robot is successful in earning the trust of human participants' based on the notable distinction between

the trust level before interaction and after the interaction. Moreover, humans *facial expressions, utterances* and *eye contact* with the robot reflected a decrease in the negative attitude of human participants after interaction with the robot. The human participants' negative attitude towards the robot (somewhat reflected by their behaviour), must have changed as human participants became more open to interacting with the robot. To date, humans have very low exposure to physically present robots in their personal life, and therefore their perception towards them is influenced by fictitious media. It is expected that as the number of opportunities for interaction with physically present robots increases, consideration will be given to our study for future robot design metrics. Therefore, findings from this research can serve to guide future work in the identification of specific robot design standards. As far as we know, our work is the first to report the results of social research on humans' attitudes toward robots and how attitudes shift towards positive attitudes after having a positive interaction with the robot who can provide explanations.

References

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

[Nothdurft *et al.*, 2014] Nothdurft, Florian and Richter, Felix and Minker, Wolfgang *Probabilistic human-computer trust handling.* In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 51–59, 2014.

[Crigger, 2004] Nancy J. Crigger. Always having to say you're sorry: an ethical response to making mistakes in professional practice. *Nursing Ethics*, 11(6):568–576, 2004.

[Shiomi, Nakagawa and Hagita, 2013] Masahiro Shiomi, Kayako Nakagawa, and Norihiro Hagita. Design of a gaze behavior at a small mistake moment for a robot. *Interaction Studies*, 14:317–328, 01 2013.

[Swartout *et al.*, 1993] Swartout, William R and Moore, Johanna D *Explanation in second generation expert systems.* In *Second generation expert systems*, 543–585, Springer, 1993

[Haring *et al.*, 2014] Haring, Kerstin Sophie and Silvera-Tawil, David and Matsumoto, Yoshio and Velonaki, Mari and Watanabe, Katsumi *Perception of an android robot in Japan and Australia: A cross-cultural comparison.* In *International conference on social robotics*, 166–175, Springer, 2014

[Dzindolet *et al.*, 2003] Dzindolet, Mary T and Peterson, Scott A and Pomranky, Regina A and Pierce, Linda G and Beck, Hall P *The role of trust in automation reliance.* In *International journal of human-computer studies*, 58(6), 697–718, Elsevier, 2003.

[Darlington and Keith, 2013] Darlington and Keith *Aspects of intelligent systems explanation.* In *Universal Journal of Control and Automation*, 1(2), 40–51, Horizon Research Publishing, 1995.

[Shaw-Garlock and Glenda, 2009] Shaw-Garlock and Glenda *Looking forward to sociable robots* In *International Journal of Social Robotics*, 1(3), 249–260, Springer, 2009.

[Schaefer and Kristin, 2013] Schaefer and Kristin *The perception and measurement of human-robot trust*, 2013

[Gaudiello *et al.*, 2016] Gaudiello, Ilaria and Zibetti, Elisabetta and Lefort, Sébastien and Chetouani, Mohamed and Ivaldi, Serena *Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers.* In *Computers in Human Behavior*, (61), 633–655, Elsevier, 2016.

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

[Chaplin and James Patrick, 1968] Chaplin and James Patrick *Dictionary of psychology*, 1968.

[Nass *et al.*, 2000] Nass, Clifford and Moon, Youngme *Machines and mindlessness: Social responses to computers.* In *Journal of social issues*, 56(1), 81–103, Wiley Online Library, 2000.

[Yagoda *et al.*, 2012] Yagoda, Rosemarie E and Gillan, Douglas J *You want me to trust a ROBOT? The development of a human–robot interaction trust scale.* In *International Journal of Social Robotics*, 4(3), 235–248, 2012.

[Heerink *et al.*, 2012] Heerink, Marcel and Kröse, Ben and Evers, Vanessa and Wielinga, Bob *Assessing acceptance of assistive social agent technology by older adults: the Almere model.* In *International Journal of Social Robotics*, 2(4), 361–375, Springer, 2010.

[Nomura *et al.*, 2006] Nomura, Tatsuya and Kanda, Takayuki and Suzuki, Tomohiro *Experimental investigation into influence of negative attitudes toward robots on human–robot interaction.* In *Journal of* Ai & Society, 20(2), 138–150, Springer, 2006.