Trust Me ! I Can be a Designated Driving Assistant

Misbah Javaid
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: misbah.javaid@griffithuni.edu.au

Vladimir Estivill-Castro
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: v.estivill@griffith.edu.au

Rene Hexel
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: r.hexel@griffith.edu.au

***Abstract*—*Autonomous vehicles* drive themselves by utilizing sensors and artificial intelligence. Evidence from surveys has shown that humans are captivated by *autonomous vehicles*, yet reluctant to give up control entirely to an *autonomous vehicle*. Inadequacy of humans trust has been identified as a pre-eminent factor behind the unacceptability of *autonomous vehicles* for driving. We propose that explanations describing behavioural decisions serve to upgrade human's sense of trust in the driving performance of *autonomous vehicles*. The contribution of our proposed research is tested by creating an interactive scenario with 34 human participants, in which we present a robot as a *driving assistant* of an *autonomous vehicle*. We incorporated the *driving assistant* with the capability to explain *traffic rules* and *traffic signs*. Moreover, the *driving assistant* is equipped with the ability to make decisions on uncertain road situations in terms of explaining, i.e., what should be a decision and why; keeping in view *traffic rules*. Additionally, the *driving assistant* has the ability to analyse and explain when to overstep a *traffic rule*, relative to a perceived hazard on the road. During the interactive scenario, the human participants performed a decision making task comprised of different *road problem-solving* scenarios with the *driving assistant*. We examined the effect of explanations from the *driving assistant* on humans' trust under two conditions (Condition 1): *no-error* and (Condition 2): *error-justification and correction*. Overall, the results show that during the decision-making task, the human participants trusted and conformed more with the *driving assistant's* decisions as compared to their own decisions. Furthermore, the human participants perceived the decisions of the *driving assistant* under Condition 1 more reliable, intelligent and trustworthy than under Condition 2. We conclude that explanations disseminating behavioural decisions are an effective communication modality that can help to improve humans trust and perceived agency (functional capability) of *autonomous vehicles*.**

Keywords–Human-Robot Trust; Explanations;Queensland Traffic Rules;Human-Robot Physical Interaction.

I. INTRODUCTION

Autonomous vehicles are the vehicles that can drive themselves using their sensors and artificial intelligence; therefore, they need no direct input from a human. The lack of control has caused fear and speculations about the reliability of *autonomous vehicles* [1]. Notably, the behaviour of autonomous vehicles is unpredictable to humans under uncertain road conditions, and makes humans think about what a vehicle will do and why? [1]. There is no doubt that autonomous vehicles predominantly offer many benefits i.e., road accidents will be reduced because, it is reported that more than 90 % of vehicle accidents involve human factors like distractions, fatigue, and misjudgement of the situation [2]. However, introducing such vehicles on the road also manifests different challenges. Evidence from investigations has shown that humans are captivated by *autonomous vehicles*, but reluctant to completely give up control entirely to *autonomous vehicles*. One of the foremost challenges for autonomous vehicles is the inadequacy of humans trust and humans have different concerns to justify their position behind the unacceptability. For example, what if a human wants to take back the control of a vehicle and the vehicle is incapable for that, or if some bad incident happens, who is going to take responsibility for the incident [1]. However, when people do have the possibility to take back the control, the quality of the take over action varies for different traffic situations [3] and this is likely to be related to a loss of situation-awareness [4]. Research has already begun to develop strategies to address these challenges and concerns. One way to give people the feeling that they are in control, while they are not actually in control, is to provide explanations. Especially, explanations that provide decision transparency, in terms of explaining a best decision based on a road condition. Moreover, explanations can also help humans track the performance and capabilities of *autonomous vehicles*.

We designed an experimental study based on an interactive scenario with 34 human participants, in which we present a robot as a *driving assistant* of an *autonomous vehicle*. The *driving assistant* is expert in identifying and explaining *traffic rules* and *traffic signs*. Moreover, the *driving assistant* has the ability to make decisions under uncertain road situations and can make some judgements to break *traffic rules* if perceive any hazard on the road. The *driving assistant* reveals the *transparency* of its decisions by generating relevant and meaningful explanations in terms of *how* a decision is made and *why* the decision is best according to the traffic situation; keeping in view traffic rules and regulations. Our main aim is to establish humans trust through explanations that will also keep the *human-in-the loop* by supplying the correct situation-awareness. The *driving assistant* provides explanations through communicating plausibly, and also provides other explanations when necessary i.e., explaining complicated terms. This strategy will not only allow the human participants to monitor the performance ability of the *driving assistant*, but also help them understand what is going on and consequently establish trust in the *driving assistant*. In general, explanations are given to impart, modify or clarify knowledge [5], to make things clear and understandable, and are often the core of any trustworthy relationship. Even in human-human interactions, unexpected and unforeseen circumstances can affect trust, and the loss of

 ISBN: 978-1-61208-761-0

trust can be reduced by giving explanations [6]. In this sense, trust and explanations seem to be common partners in everyday life.

Revising the primary purpose of our study, we created an interactive scenario, in which human participants perform a decision-making task with a *driving assistant*, in a given time. The decision-making task is based upon *road problem-solving* scenarios. We told the human participants that the task is based on the collaboration with the *driving assistant*, and the final decision does not depend on the human participants' decision solely. We explained this to every human participant. First, human participants must choose an option, and then they can change the answer after listening to the *driving assistant's* explanations or to leave it as it was. For the proposed method, we set the focus of our inquiry through humans' acceptance and conformation to the *driving assistant's* answers, as a new objective measure of the trustworthy relationship.

If *autonomous vehicles* make the right situation assessments and provide the right explanations for their decisions, they will earn humans trust more. However, *autonomous vehicles* like other autonomous systems, can have some degree of errors or can be susceptible to misjudgement of traffic situations and that may have a significant adverse effect on humans trust towards the *autonomous vehicles*. From a performance standpoint, if *autonomous vehicles* can sense their errors and recover themselves automatically, they will be considered more efficient and reliable by humans. Similarly, providing appropriate explanations can reduce the negative effect of the situation [7]. Mapping it into real life, humans tend to trust humans if they can explain to us what do they do and why? This reflects how trust appears to work; it involves (more or less elaborate) explanations of a person or a thing that we may or may not trust. If we expect others to justify their failures to us, the same we expect from *autonomous vehicles*. Therefore, we manipulated the *driving assistant* to make a wrong judgement on some traffic scenarios, and produce inaccurate information by generating wrong explanations intentionally. A wrong explanation means that the *driving assistant* contradicts a traffic situation for some reason and produces wrong explanations. However, immediately corrects himself with a sophisticated justification for the error and sets out to understand the influence of *error-justification and correction* policy. In the time-sensitive task, if the *driving assistant* recovers from a failure and justifies the cause of the failure to the human participants, does it mitigate the possible adverse effects of the erroneous situations? Will human participants agree to rebuild trust in the *driving assistant*? By addressing these questions, the current study contributes to the design metrics of future *autonomous vehicles*. One more factor to consider is explanations modality; how the explanations should be communicated to humans according to their expectations and needs. For *autonomous vehicles* to communicate explanations to the humans, they need to construct a form of agency. The agency can be ascribed based on their ability to follow the same modalities as between human human interaction so that humans perceive the *autonomous vehicles* believable and trustworthy.

Speech is the most common mode of interaction, and many studies suggest the use of verbal statements to express information [8] for the development of humans trust in automation [9]. *Text to speech* voice exerts significant effects on humans' perception and trust in technology [10]. Therefore, the current research adopts a more direct form of communication for the provision of explanations as *English like sentences* in audio modality. We divided this paper into different sections. Section II investigates the literature on trust between humans and automation. Section III discusses the design of a robot as a *driving assistant*. Section IV describes the proposed study as well as hypotheses, experiment procedure, and measures of dependent variables. Section V details the results obtained based on our hypotheses. Section VI presents the discussion and the limitations of the study. Finally, Section VII considers the implications of the study while summarizing the conclusions.

II. Human-Automation Trust

Mayer [11] investigated that trustworthiness is based on the benevolence, perceived integrity and the ability of the given system. In this manner, Lee and See [12] also proposed a dynamic process model that guided how to build trust in automation and its impact on reliance. Their conceptual model provides insightful information about trust in automation and describes guidelines that can help in calibrating trust appropriately, thereby avoiding misuse and disuse of trust. If humans do not trust autonomous systems, the interaction between humans and systems may be affected and eventually lead to the abortion of future interactions [5]. Hoff and Bashir's model [13] also described three layers of trust. According to the model, during interaction with an automated system, trust is moderated by how a system performs during the task, its design features and the experience of the interaction itself. In this way, previous experience with a system helps to build trust in autonomous systems. In today's *semi-autonomous vehicles*, trust is aimed to be established by employing interfaces that display the automated function of the vehicle to provide humans with the transparency of the internal system [14]. McKnight [15] also suggested that trust in *autonomous vehicles* can be enhanced by focusing on certain factors, one of which is system transparency.

III. Robot as Driving Assistant

Pepper is a social robot with a height of 1.2 meters and is suited for easy Human-Robot Interactions. For perception, the *Pepper* robot has two cameras with a native resolution of 640*480 pixels. We chose *Pepper* robot for our study because, given the height of the robot, the top camera is a natural choice for our intercative scenario as it points toward the average height of a human. In our research, we created different flash cards. Each flash card contains an image and a *QR code*. A *QR code* is a quick response that can store a lot of helpful information and is similar to a barcode in matrix form. Although, a *QR code* is readable from any direction, however, the detection of the *QR code* depends on the resolution of the printed *QR code*. Hence, ensure that only high resolution *QR codes* are used. We used *QR codes* as compared to the *Nao marks* because a *QR code* is detected more accurately and allow to store a large amount of data.

During the experiment, the robot's behaviour was completely autonomous. To do this, all the explanations of the robot are preprogrammed in advance and stored in each *QR code* as an identifier for each image. We created a set of three explanations for each flash card, and the robot randomly

 ISBN: 978-1-61208-761-0

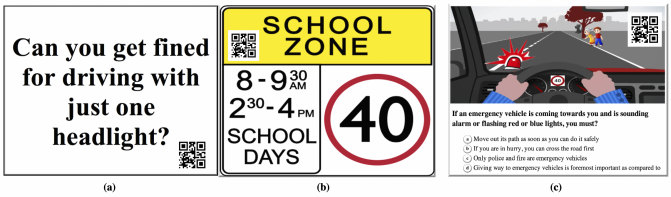

Figure 1. Images on the Flash Cards - (a) *TYPE - 1* (*Traffic Rule*), (b) *TYPE - 2* (*Traffic Sign*), (c) *TYPE - 3* (*Road Problem Solving - Hazard Perception Scenario*)

chose any explanation. In addition, if a human will show a flashcard to the robot more than once, the explanation given previously will not be triggered again. We make the human participants think that they are interacting with an intelligent *driving assistant*, who is expert enough to remember *traffic rules* and recognise *traffic signs* and can make decisions on uncertain road situations accordingly, by generating a different set of explanations every time. The robot uses the *QR code* to identify the image and to generate different explanations according to the image printed on the flash card. The robot is directed by an executable *NAOqi*, which acts as a broker and starts automatically when *NAOqi OS* starts. *NAOqi* framework contains a *ALBarcodeReader* vision module, that is used to recognise and decode a barcode. The robot uses the *ALBarcodeReader* vision module, using *Python* (an interpreted language) to scan an image in the camera and find a *QR code* in the image. If a barcode is detected in the image, the module will try to decypher it and raise an event to trigger *ALTextToSpeech* (this is another module of the *NAOqi* framework), which enables the robot to speak. We created a separate database for the images printed on the flashcards, and the explanation for each image in *MySQL*. To keep a human in the loop, we also developed a *Graphical User Interface* in *Python* for human participants, that contains images with numbers. Every time, a human participant shows the image to the robot, the experimenter selects the same image on the *Monitor Screen*, so that the human participant can see the image and listen to explanations from the robot. There is an operator who monitors the robot and can control the physical movements of the robot. Also, the operator can make the robot speak as a result of unforeseen questions from the human participants.

A. Experiment Material

We made 105 flash cards with different images on *traffic rules*, *traffic signs* and *road problem-solving* scenarios. All the images and explanations were created by the *Queensland Department of Transport and Main Roads* [16]. We created three possible types of flash cards (35 of each type). Figure 1 shows an example of each type of flash card.

1) TYPE - 1 Flash Cards with Traffic Rules): Type 1 flash cards contain only *traffic rules* and are written in text format, as shown in Figure 1 (a). The primary goal with the *Type 1* flashcard is, we want a human to observe the reading ability (correctly reading without making any mistake) and ability of the *driving assistant* to produce correct and relevant explanations according to the image on the flashcard. Expected explanation from the *driving assistant* for *Type 1* flashcard is:

"Listen human carefully ! With only one working headlight, you cannot drive at night or in conditions of low visibility. Even in the daytime, you may get pulled over if you are seen with only one headlight. If your vehicle has other faults or your headlight has been out for a while, you may be fined."

2) TYPE - 2 Flash Cards with Traffic Signs: Figure 1 (b) shows a Type 2 flash card containing only *traffic signs*. The primary goal with the *Type 2* flash cards is we want a human to analyse that the *driving assistant* is not only capable of reading, but it also has a correct assessment of the *traffic sign* and then producing relevant and explanations according to its assessment. The *driving assistant's* explanation for the Figure 1 (b) is:

"School zone speed limits do not apply on public holidays or weekends. This sign means, you must keep a speed of 40 kilometers per hour in the morning from 8 am to 9:30 am and in the afternoon from 2:30 pm to 4 pm during school days only."

3) TYPE - 3 Flash Cards with Road Problem Solving Scenarios: Type 3 flash cards contain road problem solving scenarios, which are more similar to hazard perception scenarios. Figure 1 (c) shows *Type 3* flash card. If the *driving assistant* can read and explain *traffic rules* and correctly evaluates *road signs*, it does not necessarily mean that the *driving assistant* complies with *traffic rules* and can be trusted as a *driving assistant*. Many people think and admit that the *traffic ruless* are the guidelines for human drivers to drive a vehicle on the road. So, a good driver is the one who knows when to break a traffic rule so as to keep the situation safe and to avoid any accident on the road. Keeping in view, we have created the *road problem solving* scenarios, which are more about solving problems by making decisions under uncertain road conditions and communicating those decisions to humans in terms of relevant explanations. The main goal with *Type 3* flashcards is that we want humans to be able to assess that the *driving assistant* can do something meaningful in a complex situation.

"Giving way to emergency vehicles is foremost important as compared to yourself and other road users. If an emergency vehicle is coming towards you and is sounding an alarm or showing flashing red or blue lights, you must move out its path as soon as you can do so safely. However, giving a way to emergency vehicles should always be done with the utmost care and with the safety of yourself and all other road users as a priority. For pedestrians and other road users, this is correct but, in this scenario, I am driving an ambulance at 40 km/h speed I can see an old lady with her son is about to cross the road and I also saw a dog hiding behind the tree. The dog also wants to cross the road. I am sounding the horn, but the pedestrians do not seem to respond. The situation is not safe for me so I will reduce my speed to stop because there is no traffic on the road and wait for the pedestrians to cross the road first then, I will move on. Remember Human! Emergency vehicles often stop or slow down to check if they can pass through safely."

IV. Proposed Study

To explore the effect of explanations on the human participants' trust towards the *driving assistant*, we carried out an experimental study.

 ISBN: 978-1-61208-761-0

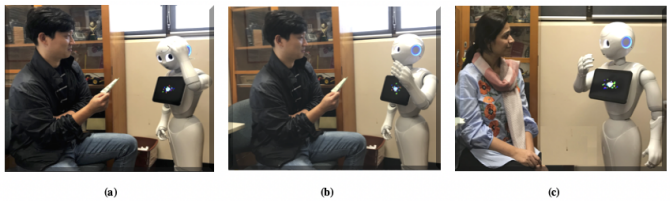

Figure 2. After a fauly behaviour, the *driving assistant* corrects himself (a) The *driving assistant* is scratching head and recalling the correct explanations (b) The human participant maintains eye contact with the *driving assistant* and is listening carefully to the explanations (c) The human participant is looking at the *driving assistant* with strange facial expressions.

A. Hypotheses

We aim to extend our line of research by posing the following hypothesis :

- *Hypothesis 1* - Humans would appreciate being informed of the *decision-transparency* of the *driving assistant* in terms of explanations and that would facilitate the establishment of trust.
- *Hypothesis 2* - Explanations that disseminate *error-justification and correction* (after the faulty behaviour), help to remedy negative effect of the erroneous situation and rebuild humans' trust in the *driving assistant*.
- *Hypothesis 3* - During a decision-making task, humans conform more with the *driving assistant's* decisions, as compared to their own decisions, when experiencing uncertainty in the environment.

B. Design of Experiment

We use a between-subjects design for our experiment, in which the human participants interact with a *driving assistant* in two possible conditions:

- *Condition 1 - control condition*: The *driving assistant* makes no mistake and provides *correct explanations*
- *Condition 2 - error-justification and correction*: The *driving assistant* makes an error intentionally and provides wrong explanations, but immediately corrects the error with a sophisticated justification.

During the decision-making task, the *driving assistant* did not directly answer by telling which option was correct. Rather, the human participants have to use their common sense to verify the correct option by listening carefully to the *driving assistant's* explanations. In *condition 2*, to justify its failure, the *driving assistant* generates different words along with gestures for example : *"I am sorry for wrong assessment and scratches head to show that it is recalling the correct explanations", "oh wait human, let me have a look again", AND "Sorry I don't agree with you human, my belief is...!"* (See Figure 2 (a), (b) and (c)). In this way, we also assessed the impact of *acceptance of mistake* from a *driving assistant* towards the human participants.

C. Humans' Conformation to the Driving Assistant as an Innovative Measure of Trust

In the field of human-robot interaction, many subjective measures of humans trust in robots have been developed and

Figure 3. The human participants are showing flash cards to the *driving assistant*.

are mostly based on self-reports (i.e., questionnaires). The 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 the human's inner belief and are limited in their capacity to analyse further on which robot knowledge, the human has built its trust. One complementary approach in this perspective is *Media Equation Theory* [17], which illustrates that, when humans engaged in collaborative tasks with computers, they tend to accept computers as social entities unconsciously. Therefore, they trust the 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, we measured human participants' conformation to the *driving assistant's* answers; generated in terms of explanations, to specific questions as an innovative measure of humans' trust in the *driving assistant*. In particular, we want to examine the humans' trust in the *driving assistant's* competency by assessing its correct situation awareness and (1) change their answers after getting explanations, (2) or reject the *driving assistant's* answers and stick with their own answer(s).

D. Procedure of Experiment

In the previous literature, human-robot trust has been measured either by objective measures (implicit) or by subjective measures (explicit). Objective measures can be retrieved from behavioural data (i.e. response time) unconsciously produced by individuals and subjective measures deals with self-reports and questionnaires retrieved from collected verbal data consciously produced by the individuals [18]. The former is limitedly developed in human robot interaction, while the latter is widely used. This study adopted the approach of combining survey(s) with an experiment to evaluate the humans' trust in the *driving assistant*. We conducted experiment in three stages.

1) Stage - 1 of Experiment: During *Stage 1*, we evaluated human participants' initial level of trust towards the *driving assistant*, by filling Human-Robot Trust questionnaire [19] as pre-interaction questionnaire.

2) Stage - 2 of Experiment: During *Stage 2*, initially, human participants selected six flash cards (three of each type i.e., *Type 1* and *Type 2*) and showed to the *driving assistant* sequentially and listened to the explanations. Following this, the human participants performed the decision-making task with the *driving assistant*, as per the following steps:

1) Human Participants selected three different flash cards of *Type 3* from a pile of flash cards.
2) Meanwhile, the *Monitor Screen* displayed the scenario, and the human participant after analysing, solved the scenario by selecting an option(s).
3) Following this, the human participant was given a chance to change its answers after listening to the

 ISBN: 978-1-61208-761-0

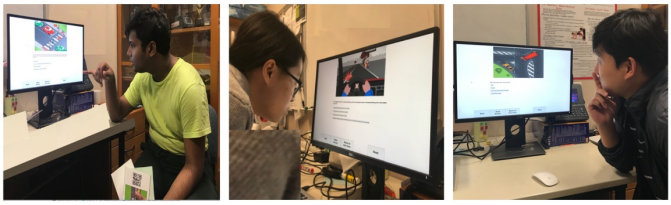

Figure 4. The human participants are selecting the correct option(s) according to the given scenario.

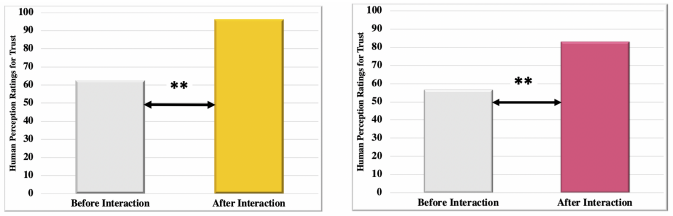

Figure 5. Difference in the trust level of human participants before and after interacting with the *driving assistant* (a) (Condition - 1) *no-error* (b) (Condition - 2) *error-justification and correction* - (**Correlation is significant at $p < 0.01$).

driving assistant's explanations, otherwise the answer was saved. For each scenario, we gave each human participant 150 seconds.

Figure 4 shows the human participants are solving the traffic scenarios.

3) Stage - 3 of Experiment: Trust is a dynamic attitude that changes over time [12] [19]. On the completion of the experiment with the *driving assistant*, as a possible clarification of the change in the humans' trust in the *driving assistant*, the human participants filled another Human-Robot Trust questionnaire [19] after interaction.

E. Measures

The independent variable was the explanations before and after interaction with the *driving assistant*. For quantitative assesment, subjective and objective analyses of the interaction were performed. The dependent variables are divided into two categories:

1) Human participants' trust, which is not directly observable, by using a 14-items subscale of the Human-Robot Trust questionnaire [19], which focuses specifically on the robot's functional capabilities, before and after interaction.
2) Impact of explanations was also analysed with the following questions:
   - do you believe the *driving assistant* "knows" the *Traffic Rules*?
   - Do you believe the *driving assistant* "follows" the *Traffic Rules*?
   - Do you trust in the *driving assistant*?

We video-recorded the experiment to examine the affective states and behavioural responses of the human participants towards the *driving assistant*, especially during the decision-making task.

F. 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 human participants in *condition 1* and 17 human participants in *condition 2* as well). We recruited human participants through general advertising, using posters on university notice board, and communicating directly with students. Each human participant received an invitation letter for the main objective of conducting the experiment. We offered a gift card valued *AUD* 10 as a *token of appreciation* to every human participant.

V. Experiment Results

In this section, we present the results of the subjective and objective assessments of the effect of explanations on human participant's level of trust, set in the context of the human-robot collaborative scenario. Before conducting any analysis, we performed a reliability analysis (Cronbach's α) to assess the internal reliability of the Human-Robot Trust questionnaire [19] and it was $\alpha > 0.723$. An $\alpha > 0.7$ or higher is considered acceptable, indicating the reliability of the measuring scales. Following this, we performed a normality analysis using *Shapiro-Wilk Test* to check whether the dependent variable trust follows a normal distribution. The test reported a normal distribution.

A. Condition - 1 : Controlled Condition (No-Error)

We performed a parametric paired sample *t-test* to analyse the overall effect of the explanations from the *driving assistant*. After interacting with the *driving assistant*, we compared the trust levels of human participants, controlling the levels of trust reported before interaction. Results showed a significant difference (t(16) = -7.512, p<0.001), suggesting that the paired sample *t-test* is appropriate in this case. Figure 5 (a) shows a glimpse of the effect of explanations from the *driving assistant*, that reflects significant higher trust levels after interaction (M = 96.41 $\pm$ 4.63), when compared with the trust levels reported before interaction (M = 62.76 $\pm$ 7.69).

B. Condition - 2 : Error-Justification and Correction

With the help of parametric paired sample *t-test*, we analysed the effect of *driving assistant's* faulty behaviour on the human participants' trust. We examined human participants' trust towards the *driving assistant* when it produced an error but corrected himself immediately with that of before interaction. The results showed a significant difference (t (16) = -22.50, p<0.001), suggesting the suitability of dependent samples *t-test*. Figure 5 (b) shows significant higher trust levels towards the *driving assistant* after interaction (M = 83.06 $\pm$ 8.52), when compared the trust levels before interaction (M = 56.63 $\pm$ 6.19).

C. Impact of Explanations by General Question Items

No matter whether the *driving assistant's* behaviour is *error-free* or it makes mistakes in predicting the behaviour of other road users; because it immediately corrects himself by selecting and implementing the most appropriate response, therefore, it can help a human to drive. The human participants realised that the *driving assistant* was competent in detecting

 ISBN: 978-1-61208-761-0

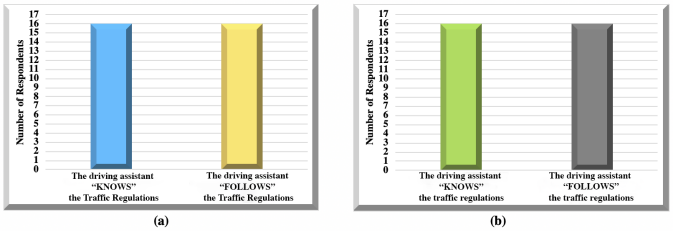

Figure 6. (a) Explanations with *no-error* (b) explanations with *error-justification and correction*.

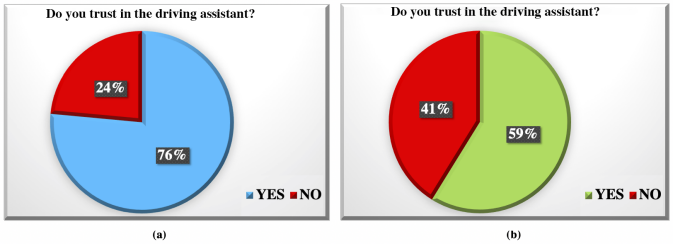

Figure 7. (a) Explanations with *no-error* (b) explanations with *error-justification and correction* policy.

hazards, and also explained the right decision according to traffic rules and regulations. Even if it considered to break traffic rules, it was only to minimise the likelihood of an accident. Therefore, the *driving assistant* "knows" the traffic rules and "follows" the traffic rules (refer to Figure 7 (a) and (b)), it can be trusted to help humans to drive safely (refer to Figure 7). However, in general, the human participants who received explanations without errors trusted in the *driving assistant's* ability more.

D. Human Participants Conformation to the Driving Assistant

In addition, we also kept a record of the number of times the human participant changed an answer after listening to the *driving assistant's* explanations. If the human participant changed the answer after explanations, then we can say that the human trusted the functional capabilities of the *driving assistant*. Our method to calculate the conformation score was to divide the number of times a human participant changed its answer to the *driving assistant's* answer by the total number of times where the *driving assistant's* answer mismatched with the human participant's answer selected for the first time. Therefore, we got a reasonable score for the analysis ranging between 0 (no conformation) and 1 (full conformation). A score greater than or equal to 0.5 was considered as human participant's trust in the *driving assistant*, see Figure 8 for conformation score. Interestingly, human participants were willing to accept and conformed more to the *driving assistant's* answers as compared to their answers. To examine whether a group of human participants under *Condition 1* conformed more with the *driving assistant* or a 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. Hence, we performed (non-parametric) *Mann-Whitney U Test* for paired samples, which indicated no significant difference between the

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

Figure 8. Conformation score for the group of human participants, N = 17 in each group.

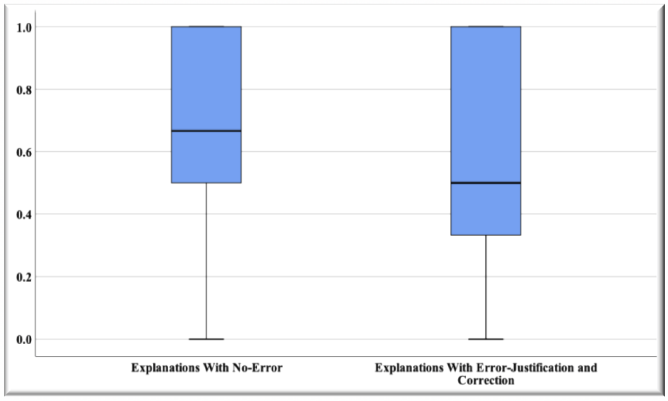

Figure 9. Human participants' conformation with the *driving assistant's* decisions.

Figure 10. After selecting the option(s), the human participants are verifying their selected option by asking from the *driving assistant*.

two groups (Z = -0.090, p>0.05), as shown in Figure 9.

Maybe, the human participants relied more on the driving assistant's decisions, because they observed that it has some criteria or logical demonstration to apply knowledge of traffic rules rather than applying blindly. Furthermore, if the driving assistant considers to break a traffic rule, it is based on an evaluation of the danger of the situation. Some human participants identified the correct option(s), still they verified by asking from the driving assistant as shown in Figure 10.

VI. Discussion

This paper conducted an experimental study to investigate whether a driving assistant characterized by the capability of providing explanations can earn the humans' trust. Specifically, we examined the effect of explanations under two conditions, i.e., (1) *no-error*, and (2) *error-justification and correction* policy. Overall, the human participants trusted the driving assistant in both conditions by supporting our *Hypothesis 1* and *Hypothesis 2*. However, the human participants under *Condition 1*, perceived the decisions of the *driving assistant* more intelligent and trustworthy. Figure 5 (a) visualizes a higher level of trust in the driving assistant. These findings motivate the acceptability of the autonomous vehicles in the human environment. By adding an extra layer of communication in terms of explanations in the design metrics of the autonomous vehicle can promote humans' trust towards them.

 ISBN: 978-1-61208-761-0

Especially, explanations that not only describe "what" should be a decision according to a traffic situation but also justify the decision by providing a sophisticated reason i.e., why is the decision best. As in our study, the humans' not only trusted the explanations given by the driving assistant but conformed more with it, supporting our *Hypothesis 3*. The driving assistant's explanations helped the human participants to scrutinize the information provided by him. The human participants have a fair understanding that the *driving assistant* has not only reasonable understanding of road rules but also has excellent ability to spot a hazard by visual scanning and detecting road-surface-based hazards. Furthermore, the driving assistant also prepares to respond i.e., to break a traffic rule to ensure that the situation is safe. Hence, humans understand and recognise the capability of the driving assistant as an expert, which was reflected during the decision-making task, the human participants' withdraw their answers and conform more to the driving assistant. This constitutes a pertinent measure to straightforwardly registering the human participants' trust in the *driving assistant*.

In addition, the strategy of *error-justification and correction* for autonomous vehicles can make humans comfortable in accepting mistakes made by it, if the consequences are not very severe. On the other hand, the strategy can also alert humans that they have to be attentive and aware of the surroundings because over trust can cause less visual attention of the road and also leads to slower reaction times when humans need to intervene in a case of an emergency. Most importantly, reactions to take-over control of the vehicle can also lead to low-quality decisions [20]. Therefore, explanations communicated in audio-modality can potentially help by keeping humans in the loop of driving assistant's decision-making process, thereby potentially avoiding a reduction in reaction times. We also noticed the human participants, after looking at the monitor screen, also scrutinize the flashcard to examine the image and to inspect whether the *driving assistant's* explanations are aligned with the image on the flashcards as shown in Figure 11. The human participants aimed at assessing the trust in the *driving assistant's* functional capabilities by considering it safe who knows how to recognise and respond to hazards. We also analysed the voices of the human participants, especially under *Condition 2* as wao, genius, intelligent and maintained an eye contact with the driving assistant.

In the end, we gave the human participants a chance to give their free opinion, and many of them wrote different comments for the performance and ability of the *driving assistant*: *"Such driving assistants can help people to follow the rules", "Although I said the driving assistant can be trusted to drive but it cannot be fully trusted because it makes mistakes and then corrects himself, but again that is a good thing", "The driving assistant not only knows the rules but it also knows how to apply the rules, which is surprise to know that it can perform so much", "Robots may not be able to make rapid decisions on empathy, but it makes decisions on facts and rules only", "I do support autonomous vehicles."*

A. Limitations

The current study used an interactive scenario with an autonomous *driving assistant*, to investigate the effects of explanations in improving humans' perceived agency (functional capability) and trust in *autonomous vehicles*. Although

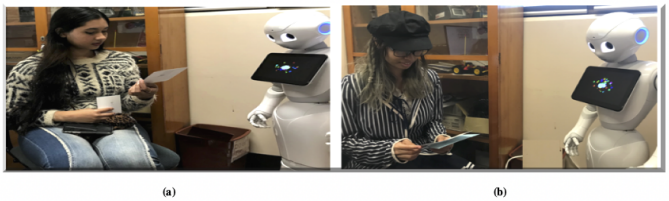

Figure 11. (a) and (b) The robot is giving explanations and the human participants are looking into flash cards to scrutinize whether the *driving assistant's* explanations are aligned with the image on the flash card.

the interactive scenario allowed us to perform experimental control, it does not have sustainability in real traffic situations. There is a considerable difference between a stationary autonomous *driving assistant* that is prone to make little errors in making judgements of the traffic situations with that of an *autonomous vehicle* that makes errors in real road traffic situations. When such *autonomous vehicles* share roads with humans, the limit will become obvious. Hence, the perception and trust of the human participants in the *driving assistant* has limited impact without any danger. In our daily lives, not every situation requires explanations, and in most cases humans mainly need explanations for circumstances that do not meet their expectations. The same is true for *autonomous systems*; humans often need explanations for autonomous decisions, which can confuse them. For an *autonomous vehicle*, if it is always *error-free* and behaves as expected, there might be no need for explanations. This seems to be compatible with the trend in our results and the choice of our experimental study. The explanations in the study were simulated through *text-to-speech* commands along with necnecessary pauses, to create a natural tone in the voice of the *driving assistant* and enough to create a significant impact on the humans, which has been demonstrated by the results of this research. We expect if *autonomous vehicles* behave intelligently by understanding, which situations to be explained. This will contribute to upgrading humans trust.

VII. Conclusion

This main purpose of this study was to investigate the effects of explanations from a *driving assistant* in the level of human participants. In this perspective, we implemented an interactive scenario in which we presented a robot as an intelligent *driving assistant* of an *autonomous vehicle*. We enhanced the capability of the *driving assistant* to enable it to recognise and explain traffic rules and traffic signs. Moreover, the *driving assistant* has the ability to solve road problem solving scenarios by making decisions in uncertain road situations and is competent enough to break a traffic rule to minimise the likelihood of an accident.

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 *driving assistant* is successful in earning the trust of human participants'. The appearance of fully *autonomous vehicles* on the roads seems to be very close. To date, humans have very low exposure to physically present *autonomous vehicles*, so their perception has been shaped by fictitious media. We expect that as the opportunity for interaction with

real *autonomous vehicle* increases, findings from the study can serve to guide future work in the identification of specific *autonomous vehicles'* design standards. This research has the potential to promote the acceptability of *autonomous vehicles* in human environment by addressing the topic of trust through explanations.

REFERENCES

[1] B. Schoettle and M. Sivak, "A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia," University of Michigan, Ann Arbor, Transportation Research Institute, Tech. Rep., 2014.

[2] M. Peden, "The world report on road traffic injury prevention: getting public health to do more," Journal of Geneva, Switzerland: World Health Organization, 2005.

[3] Radlmayr et al., "How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving," in Proceedings of the human factors and ergonomics society annual meeting, vol. 58, no. 1. Sage Publications Sage CA: Los Angeles, CA, 2014, pp. 2063–2067.

[4] D. Winter et al., "Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence," Transportation research part F: traffic psychology and behaviour, vol. 27, 2014, pp. 196–217.

[5] F. Nothdurft, F. Richter, and W. Minker, "Probabilistic human-computer trust handling." in SIGDIAL Conference, 2014, pp. 51–59.

[6] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in Proceedings of the 13th international conference on Intelligent user interfaces. ACM, 2008, pp. 227–236.

[7] Khastgir et al., "Calibrating trust to increase the use of automated systems in a vehicle," in Advances in Human Aspects of Transportation. Springer, 2017, pp. 535–546.

[8] Visschers et al., "Probability information in risk communication: a review of the research literature," Risk Analysis: An International Journal, vol. 29, no. 2, 2009, pp. 267–287.

[9] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in International Conference on Social Robotics. Springer, 2015, pp. 574–583.

[10] L. Qiu and I. Benbasat, "Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars," International journal of human-computer interaction, vol. 19, no. 1, 2005, pp. 75–94.

[11] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," Academy of management review, vol. 20, no. 3, 1995, pp. 709–734.

[12] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 46, no. 1, 2004, pp. 50–80.

[13] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," Human factors, vol. 57, no. 3, 2015, pp. 407–434.

[14] P. Pu and L. Chen, "Trust building with explanation interfaces," in Proceedings of the 11th international conference on Intelligent user interfaces. ACM, 2006, pp. 93–100.

[15] D. H. McKnight and N. L. Chervany, "Trust and distrust definitions: One bite at a time," in Trust in Cyber-societies. Springer, 2001, pp. 27–54.

[16] Department of transport and main roads. [Online]. Available: https://www.tmr.qld.gov.au [retrieved: March, 2020]

[17] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," Journal of social issues, vol. 56, no. 1, 2000, pp. 81–103.

[18] Gaudiello et al., "Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers," Computers in Human Behavior, vol. 61, 2016, pp. 633–655.

[19] K. E. Schaefer, "The perception and measurement of human-robot trust," Ph.D. dissertation, University of Central Florida Orlando, Florida, 2013.

[20] Gold et al., "Take over! How long does it take to get the driver back into the loop?" in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 57, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2013, pp. 1938–1942.

 ISBN: 978-1-61208-761-0