Towards Anthropomorphising Autonomous Vehicles: Speech and Embodiment on Trust and Blame After an Accident

1st Christopher D. Wallbridge, 2nd Victoria Marcinkiewicz, 3rd Qiyuan Zhang, 4th Phil Morgan
IROHMS and HuFEx, School of Psychology
Cardiff University
Cardiff, UK
wallbridgec@cardiff.ac.uk, marcinkiewiczv@cardiff.ac.uk, zhangq47@cardiff.ac.uk, morganphil@cardiff.ac.uk

***Abstract*—A novel experiment is presented of our research on the effects of anthropomorphism on trust and blame after an accident involving an Autonomous Vehicle (AV). We presented 147 –out of an expected 300 based on power calculations– participants simulation software generated animations of a hypothetical accident involving an AV, with manipulation of presence of a humanoid robot, and different conversation styles. So far we have found no direct effect on trust, but we have found promising results on factors that correlate with trust; measures of competence and discomfort, and a potential effect on blame.**

***Index Terms*—Human-Robot Interaction, Self Driving Vehicles, Blame, Trust, Human Factors**

I. Introduction

Currently most road traffic accidents are considered to be caused by human error. Autonomous Vehicles (AVs) bring the promise of alleviating many of the causes of accidents e.g. fatigue, lapses in attention and ignoring rules of the road.

Six levels of automation are proposed by the Society of Automotive Engineers [1]. Here we focus on L5; the car can drive itself in all locations and under all conditions. L5 AVs are still considered a long way off. Even when they are deployed, accidents will likely still be inevitable due to complications of a real world environment [2]. This despite the potential for the technology to be safer compared to human drivers.

Accidents, or even near misses, while using an AV are likley to cause a loss of trust. This may affect the likelihood that people will adopt or continue to use AVs [3], [4]. By understanding factors surrounding the loss of trust and how blame is assigned in the aftermath of an accident, we can mitigate loss of trust or restore it –which is important if the promise of improved safety for road users can be realised.

Our own previous research found that human drivers and AVs are blamed differently for the same incident or accident [5], even when the actions and consequences are the same. While for most cases this led to the AV having higher ratings of blame, this was not always the case. Further investigation showed in part this was to do with the strength of causal cues [6], suggesting that a contributing factor is the expectations of the capabilities of human drivers vs AVs.

One potential method of exploring the dynamics of this trust is to try to anthropomorphise the AV. Current research into Human-Robot Interaction (HRI) suggests that increased anthropomorphism can lead to increased trust [7]. Within AVs, anthropomorphism has even been shown to increase perceptions of competence [8]. Further research has shown that the context is important as to whether anthropomorphism can increase trust [9]. As expectations of capability may play a role in how such systems are trusted or blamed after an accident, it is important to explore these aspects in this context.

In this paper we present early results for a study on anthropomorphism in AVs after an accident has occurred. The data presented here is only a portion of the data we are in the process of collecting. We also discuss our findings based on these initial results.

II. Methodology

A. Design

To operationalise anthropomorphism we use a 3 (Conversation Style) x 2 (Presence of Humanoid Robot) between subject study design (based on work by [8]). The levels for presence of robot were: present or not present. Conversation style, had three levels:

- *No Speech*
- *Informational* - Programmed speech would give details on the intention and actions of the AV. The speech would refer to the AV in the third person.
- *Conversational* - Programmed speech would provide the same information as in the informational style. However it would refer to the AV in the first person, and name itself.

As dependent variables we measured warmth, competence and discomfort using the the Robotic Social Attributes Scale (RoSAS) [10]. We measured trust using two more scales,

The research was funded through the ESRC Project ES/T007079/1 Rule of Law in the Age of AI: Principles of Distributive Liability for Multi-Agent Societies and is part of a larger project on the same topic supported by ESRC (ES/T007079/1) and JST. This work was also conducted with support of the Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS) operation C82092 and partially funded by the European Regional Development Fund (ERDF) through the Welsh Government.

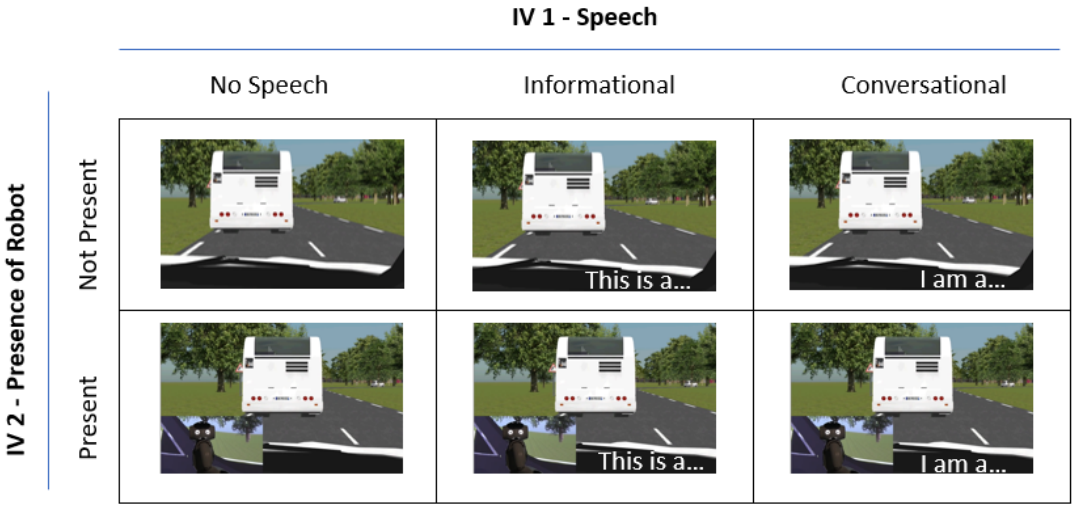

Fig. 1. Figure showing the 3x2 study design of our study, with screenshots of how the videos would look to participants, and examples of the words heard.

the Trust in Automated Systems Survey (TiAS) [11] and the Situational Trust Scale for Autonomous Driving (STS-AD) [12]. Finally we measured blame on both the AV and 3rd parties using questions based on [5]. As we are using an online survey tool we were able to adapt all likert scales to be VAS, to give us values from 0 to 100.

Our hypotheses were:

- H1 As levels of speech increased (from no speech to conversational) we would see increased trust.
- H2 Presence of the humanoid would increase trust.
- H3 We would see an interaction between presence of the robot and conversation style on blame.

B. Materials

Based on the methodology proposed in [13], we use a Simulation-Software-Generated Animation (SSGA) to create the videos of our accidents. We chose this method based on L5 AVs not being available, but also the ethical implications of attempting to expose participants to a real (or seemingly real) accident. It would also allow us to capture a much larger sample more easily at the early stages of this research.

Participants were presented with 1 of 6 videos. Every video was 1 minute and 40 seconds long. Each video would start with an introduction phase with the text: “Introduction - You are about to be shown a driving scenario already in progress”. The background would show a view of the car, as if from the driver seat you were looking towards the dashboard. After the introduction every video showed the target vehicle from the perspective of –what would normally be for a UK car– the driver’s seat. The car would follow a bus, initially on a country lane but transitioning to a town, where the bus would stop. Until this point there is a steady stream of oncoming traffic in the other lane. With the bus stopped the car starts to overtake. As the car approaches the front of the bus a pedestrian steps out in front of the car. At this point the video freezes, and informs the participant that the car was unable to stop in time and that the pedestrian suffered minor injuries.

For the speech we used the Python library for Google Text-to-speech (gTTS 2.2.4). The full script for the informational and conversational conditions can be found in table I. At all stages the scripts were designed to contain the same information, but only vary how it was presented.

If the robot was present the introduction would have a Softbank NAO v6 robot facing towards the camera. It would animate using the ‘animated say’ box provided in the Choregraphe software. After the introduction the side view, with the robot kneeling down, would be presented in the bottom left corner as a picture in picture view. An example of the view can be seen in figure 1. Whenever the robot spoke, it would turn it’s head to look at the camera, before looking back. The audio of the robot’s servo movements was also included.

As the study required participants to be able to hear the speech, and we were unable to directly check that participants had setup their sound properly we introduced a sound test. Our sound test was based on work by [14] and [15]. The intention behind our sound screening was to ensure that participants had adjusted their volume so that they could understand the words being said. Therefore our sound test used words. We created five sound files. For each sound file five words were randomly selected from the phonetic alphabet, then one word was randomly selected to be quieter than the rest. The volume of the ‘quiet’ word was set at the level of the speech that the videos would use. This was 50% of the sound for the original file (note this was using iMovie) whereas the louder words were set at 90% volume. This level of difference seemed necessary for people to consistently identify the correct words, based on a sequence of tests with co-workers, friends and family. Participants were required to select the ‘quiet’ word in 3 out of the 5 tests to proceed with the study.

C. Procedure

Participants were recruited through Prolific, where they were told this would be a study on road safety. They were

TABLE I
TABLE SHOWING THE SCRIPT THAT WAS USED FOR THE SPEECH IN DIFFERENT CONDITIONS, ALONG WITH TIMESTAMPS.

Timestamp	Informational	Conversational
0:00	This is an Autonomous Vehicle Informational Assistant, or Avia for short. You are about to be shown a driving scenario already in progress.	I am an Autonomous Vehicle Information Assistant, but you can call me Avia. You are about to be shown a driving scenario already in progress.
0:12	Vehicle is driving behind a bus on a country lane, looking for opportunities to overtake.	We are driving behind a bus on a country lane, I am looking for opportunities to overtake.
0:25	The traffic conditions are preventing vehicle finding an appropriate overtaking window.	The traffic conditions are preventing me from finding an appropriate overtaking window
0:36	The high traffic density is still preventing vehicle from overtaking.	The high traffic density is still preventing me from overtaking.
0:54	Vehicle is still prevented from overtaking.	I am still being prevented from overtaking.
1:14	The bus is stopping, providing this vehicle an opportunity to overtake.	The bus is stopping, providing me an opportunity to overtake.
1:25	Warning!	Warning!

also informed of the screening requirements, particularly that they would need to pass a sound test. This study was presented to participants as a Qualtrics survey.

Participants were first presented with an information sheet, that provided more detail on the study, what they would be expected to do, and details on right to withdraw and data policy. Each participant was then shown a consent form, again detailing their rights and data policy and that by consenting to participate that they had read and understood.

Upon consenting participants were presented with information on the sound test, which was then followed by the sound test itself. The sound test is described in section II-B. Failure to successfully complete the sound test resulted in participants being withdrawn from the study.

Participants were then asked demographic questions –age and gender– followed by whether they had a driving license, and if so, details about their driving habits. Participants were also asked a pre-trust question on AVs based on [5]: "Imagine that fully autonomous vehicles will be deployed on a large scale on UK roads within the next 12-months. Please rate how likely you would be to use an autonomous vehicle."

Further instructions to the participants, asked them to pay attention to the video, and not to try to pause, skip or repeat the video. At this point, participants were shown one of the six videos described in II-B. Having viewed the video, participants were asked a simple question on trust in the system and questions on blame on the AV, pedestrian and bus driver. The participants were then presented questions from TiAS or STS-AD, followed by the other, so that participants would answer questions from both scales. Next, participants were asked questions from RoSAS in the categories of competence, warmth and discomfort. Finally the pre-trust question was repeated. Participants were then given an opportunity to leave comments before being debriefed.

D. Participants

Currently we have collected data from 147 UK participants of an expected 300 participants. This is based on a G-Power calculation determining that we would need at least 269 participants to detect a medium effect size (Cohen's F = 0.25) with 0.8 power. We also expect to collect a matched sample from Japan to make a cross-cultural comparison at a later date.

Through Prolific, participants were screened as residents of the UK, aged $\geq$ 18, having normal or normal-corrected vision and hearing, were fluent in English and were using a laptop/desktop to complete the study. Participants were also required to have at least a 95% approval rating on Prolific. An initial 30 places on the study were released as a pilot to check our data collection. The remaining positions on the study were released over the following week.

III. RESULTS

Please note that these results are based on about half the amount of data we intend to collect based on our required power calculations. Therefore most of the results presented here are tentative. Please note unless otherwise stated that all results are based on a two-way ANOVA.

For a general overview of the results we combined the values in the TiAS. Scores were aggregated from all 12 items, with those related to distrust given a negative sign. This gives a value between -500 and 700. We have so far found no overall significant effect of conversation style (f = 0.799, p = 0.452), presence of robot (f = 0.281, p = 0.597) or the interaction between them(f = 1.648, p = 0.196). Similarly, we coded the values of the STS-AD questions together (1,3 and 6 were coded positively, the rest negatively) to give values between -300 and 300. This also presented no significance from conversation style (f = 0.573, p = 0.565), presence of robot (f = 0.007, p = 0.933) or the interaction between them(f = 0.953, p = 0.388). Overall values of trust in automation seemed mixed having seen the video (TiAS Combined mean = 16.707, sd = 210.791) and trust in the AV was low (STS-AD Combined mean = -115.1361, sd = 135.2068). Please see figure 2 for a breakdown of trust by condition in both scales.

The overall value for competence provides us with an almost significant effect for conversation style (f = 2.930, p = 0.057, Cohen's f = 0.20, means: conversational = 268.429, informational = 206.9796, none = 223.1633). Neither presence of the robot (f = 0.749, p = 0.388) or the interaction between conversation style and presence of the robot showed any significance on competence (f = 0.800, p = 0.451). When looking at values for warmth we see a significant effect for conversation style (f = 5.764, $p < 0.01$, Cohen's f = 0.29 means: conversational = 88.877, informational = 49.122, none

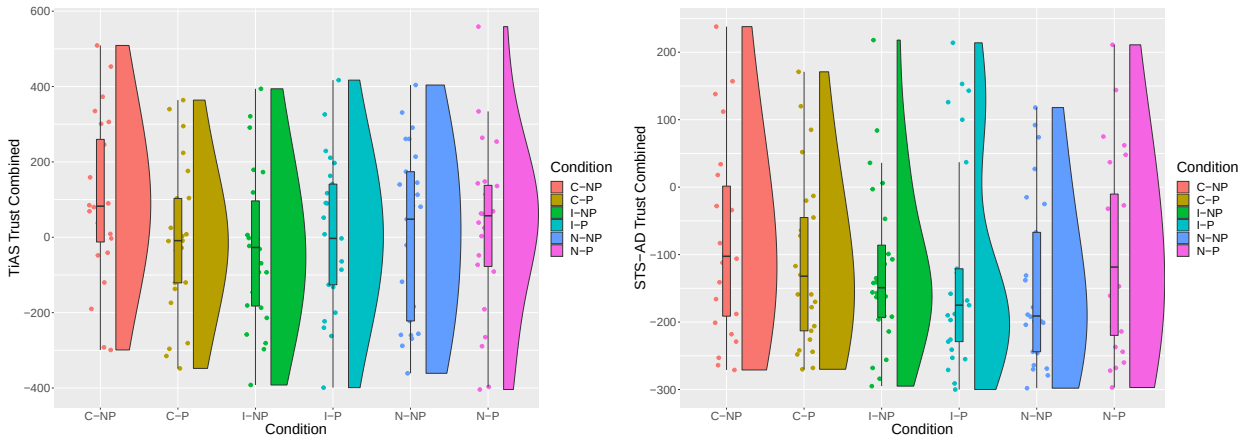

Fig. 2. Figures showing the combined values of trust for the Trust in Automation Scale (TiAS) on the left and Situational Trust Scale in Automated Driving (STS-AD) on the right. Condition is represented as ⟨conversation style⟩-⟨Presence of Robot⟩. **C:** Conversational, **I:** Informational, **N:** No speech, **NP:** Not Present, **P:** Present.

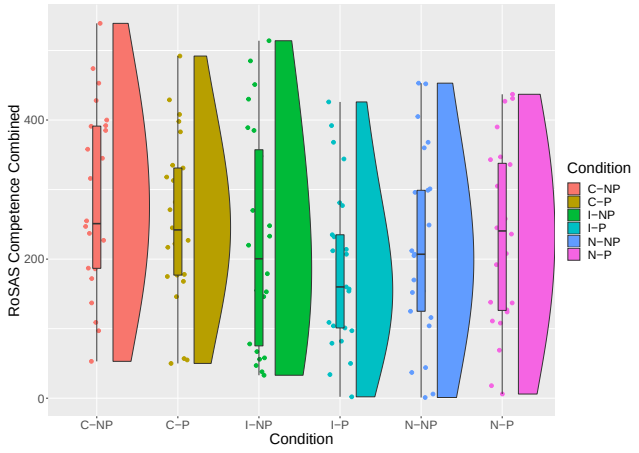

Fig. 3. Figure showing the combined value of competence from the Robotic Social Attributes Scale (RoSAS). Condition is represented as ⟨conversation style⟩-⟨Presence of Robot⟩. **C:** Conversational, **I:** Informational, **N:** No speech, **NP:** Not Present, **P:** Present.

= 44.857), but no significance for presence of the robot (f = 1.588, p = 0.210) or any interaction between the two (f = 2.135, p = 0.122). Discomfort showed a similar pattern, with a significant effect for conversation style (f = 3.623, p = 0.029, Cohen's f = 0.23, means: conversational = 157.225, informational = 205.714, none = 228.306), but with no significance for presence of the robot (f = 0.096, p = 0.758) or their interaction (f = 0.100, p = 0.905).

When we look at blame on the AV we find no main effect of conversation style (f = 1.897, p = 0.1539) or of presence of the robot (f = 0.075, p = 0.7846). For the interaction between conversation style and presence of the robot we see an almost significant effect (f = 2.530, p = 0.0832, Cohen's f = 0.19).

We found correlations (using Pearson's Product Moment) between competence and both TiAS measures of Trust (coef = 0.725, $p < 0.01$) and STS-AD (coef = 0.703, $p < 0.01$). We found a negative correlation between discomfort and TiAS measures of trust (coef = -0.547, $p < 0.01$) and STS-AD (coef = -0.527, $p < 0.01$). We also found a negative correlation between blame on the AV and both TiAS (coef = -0.490, $p < 0.01$) and STS-AD (coef = -0.687, $p < 0.01$)

IV. Discussion

Overall the values of trust are low, though likely due to the context of the accident that we are asking participants to give their ratings of trust in. Despite predictions, conversation style and presence of robot have not had an effect so far on trust after an accident. This is perhaps surprising though given the effect we see on discomfort, with increased levels of conversation style reducing discomfort. Given the medium negative correlation between discomfort and trust we might expect trust to be affected as an outcome of this. We also see a similar correlation with blame, and a stronger correlation with competence. Currently the effects of our independent variables are not yet significant, but it seems likely that with the full data collected that they will be. We cannot say at this stage if this will in turn affect trust. However this would seem to show that we should be able to affect trust, especially by the way that an AV talks to its passengers, but further research may be necessary to see how we can strengthen this effect.

With our initial 147 participants, no effect is caused by presence of the robot on trust, or our other measures. This may be caused by one of the main limitations of our study: The robot is another part of the video, rather than being present with the participant. Future work should focus on an experience with participants in a simulator where the robot is present. While this may have weakened the effects we want to see, it would not have been practical to get the number of participants we have in this study –and expect to still recruit– and put them in a driving simulator. This study should help inform a selection of the most pertinent conditions to take forward. An in person study may also allow us to look at measures of trust that are not self-reported, such as eye tracking and other bio metrics.

REFERENCES

[1] S. International. (2021) Sae levels of driving automation™ refined for clarity and international audience. [Online]. Available: https://www.sae.org/blog/sae-j3016-update

[2] P. A. Hancock, "Some pitfalls in the promises of automated and autonomous vehicles," *Ergonomics*, vol. 62, no. 4, pp. 479–495, 2019.

[3] P. H. Kim, K. T. Dirks, and C. D. Cooper, "The repair of trust: A dynamic bilateral perspective and multilevel conceptualization," *Academy of Management Review*, vol. 34, no. 3, pp. 401–422, 2009.

[4] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[5] Q. Zhang, C. D. Wallbridge, D. M. Jones, and P. Morgan, "The blame game: Double standards apply to autonomous vehicle accidents," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2021, pp. 308–314.

[6] ——, "Judgements of autonomous vehicle capability determine attribution of blame in road traffic accidents," *Submitted to Transportation Research Part A: Policy and Practice*, 2022.

[7] J.-G. Lee, K. J. Kim, S. Lee, and D.-H. Shin, "Can autonomous vehicles be safe and trustworthy? effects of appearance and autonomy of unmanned driving systems," *International Journal of Human-Computer Interaction*, vol. 31, no. 10, pp. 682–691, 2015.

[8] S. C. Lee, H. Sanghavi, S. Ko, and M. Jeon, "Autonomous driving with an agent: Speech style and embodiment," in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 2019, pp. 209–214.

[9] L. Onnasch and C. L. Hildebrandt, "Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 1, pp. 1–24, 2021.

[10] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

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

[12] B. E. Holthausen, P. Wintersberger, B. N. Walker, and A. Riener, "Situational trust scale for automated driving (sts-ad): Development and initial validation," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2020, pp. 40–47.

[13] Q. Zhang, C. D. Wallbridge, P. Morgan, and D. M. Jones, "Using simulation-software-generated animations to investigate," in *In Print in the Proceedings of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022).*, 2022.

[14] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[15] E. M. Elliott, R. Bell, S. Gorin, N. Robinson, and J. E. Marsh, "Auditory distraction can be studied online! a direct comparison between in-person and online experimentation," *Journal of Cognitive Psychology*, vol. 34, no. 3, pp. 307–324, 2022.