An Optimized Paradigm to Measure Effects of Anthropomorphized Self-Driving Cars on Trust and Blame Following an Accident*

Phillip L Morgan[abcd], Victoria E K Marcinkiewicz[abc], Qiyuan Zhang[abc], Theodor R W Kozlowski[abc]
Louise Bowen[abc], Christopher D Wallbridge[ab]

[a]Cardiff University Centre for AI, Robotics and Human-Machine Systems (IROHMS); [b]Cardiff University Human Factors Excellence Research Group (HuFEx); School of Psychology, Cardiff University, 70 Park Place, Cardiff, CF10 3AT, UK; [c]Cardiff University Digital Transformation Innovation Institute; [d]Luleå University of Technology, Psychology, Division of Health, Medicine and Rehabilitation, Sweden.

Abstract— **Despite increasing sophistication of automated technology within self-driving cars (SDCs), there have and will be instances where accidents occur. Trust could be eroded – with consequences for adoption and continued usage. At RO-MAN 2022 – we presented a SDC experiment focused on trust and blame in the event of an accident situation. We developed a novel method to investigate whether a humanoid robot informational assistant communicating SDC intentions and actions improved trust and reduced blame in such situations. One limitation was that the accident occurred with limited experience of the SDC performing maneuvers without incident. We have further developed the paradigm to include successful maneuvers to give important opportunities to build trust in the novel technology before the critical event. Initial data is presented and discussed.**

I. Introduction

The technology required to make self-driving cars (SDCs) a reality is developing at a hurtling pace: now with six defined levels of driving automation [1]. Here, we are interested in level 4-5 SDCs that can self-drive under most or all conditions that are already being deployed in some parts of the world.

Despite such technological advancements, there are many crucial unanswered questions. Arguably, a large percentage of people do not yet trust such technology to a level where they would accept and adopt it [2,3,4]. Some suggest that experience with the technology will be a key enabler of adoption [5]. That said, is mere experience enough? Given that humans are expected to be mostly out of the driving loop, should SDC actions and intentions be communicated to them during journeys, and, will this impact trust? What about when something goes wrong – leading to an accident [6]? Accidents involving SDCs have and will continue to occur and we need to better understand how best to support trust in the technology when this occurs – especially when the SDC is not at fault.

Recently, we proposed a novel method to investigate whether the level of anthropomorphism in an SDCs human-machine interface (HMI) could be beneficial in the event of an accident. A humanoid robot informational assistant was used; perceived to be part of the SDC. Limited evidence was found for increased trust and reduced blame [7].

More positive findings have been reported in contexts where SDCs perform optimally–without incident [8,9]. One study provided evidence of increased trust in an SDC when a Nao robot provided dialogue about a successful overtaking maneuver in a social ('small-talk') style vs a voice (non-conversational) only condition. However, there is a dearth of research and understanding surrounding the potential benefits of informational assistants (including robots) in situations involving SDC accidents. In some other contexts, anthropomorphism does not always promote trust in robots [10].

One key limitation of the paradigm we proposed at RO-MAN 2022 was that participants had little experience of the SDC performing optimally before the critical incident. The incident happened immediately after the SDC committed to an overtake maneuver with no previous experience of the vehicle successfully negotiating this operation.

The paradigm has been further developed in part with the SDC successfully negotiating multiple overtake maneuvers before the critical incident. The critical incident (involving a pedestrian violating the UK Highway Code) occurs at the end of the entire scenario. The informational assistant (robot or non-robot HMI) provides dialogue (conversational or informational) throughout the scenario about intentions and road conditions. In addition, the SDC within the current experiment does not perform as assertively as in our previous experiment [7] – instead – it communicates intention to overtake another vehicle only when that vehicle is stationary and it is deemed completely safe to do so.

It is predicted that trust will be higher, and blame lower with a robot informational assistant – markedly in the conversational dialogue condition.

II. Methodology

A. Participants

Well powered experiments are being conducted to detect at least medium effect sizes (Cohen's $f = 0.25$) with power of 0.8.

B. Materials, Design and Procedure

Zhang et al [11] stressed that SDC research faces a colossal methodological challenge: collecting data from sufficient samples across multiple experiments and exploring as many conditions as practicable is almost impossible. If researchers are to generate enough data and evidence from experiments with human participants to inform design recommendations, standards and regulation of SDC technology - alternative methodologies are needed. In-person experiments are

*The work was funded as part of an ESRC-JST (Economic & Social Research Council - Japan Science & Technology Agency) project ES/T007079/1: Rule of Law in the Age of AI: Principles of Distributive Liability for Multi-Agent Societies. *The first and corresponding author is the UK Principal Investigator.

important–but we can also glean important insights from high-fidelity *Simulation-Software-Generated Animations* (*SSGAs*).

Scenario. SSGAs were created using driving simulation software by SCANeR© within a bespoke AV Simulation Driving Simulator. The SSGA depicts a futuristic scenario with an in-vehicle-passenger-looking-out view of a SDC driving along a single carriage road at 30-mph/48.28-kph, first within the countryside before entering an urban town (Figure 1). Early on, the SDC approaches a moving bus and drives behind it at the speed limit and at a safe distance. The bus comes to stop at a bus-stop. Event 1 involves the SDC determining that conditions are safe (UK Highway Code) to commit to overtaking the bus: there is a broken white line separating the two lanes, no evidence of oncoming traffic in the opposite lane, and no pedestrian(s) legally attempting to cross the road. The SDC HMI (robot in some conditions, non-robot interface in others) communicates that conditions are safe to overtake, and the maneuver takes place. Event 3 is similar to Event 1. Events 2 and 4 involve the vehicle detecting traffic in the opposite lane and determining that it is too risky to overtake: it stays in the current lane and waits for the bus to move again. Event 5 is similar to Events 1 and 3. However, as the SDC drives past the bus during the overtake, a pedestrian steps out in front of the bus (critical event) violating the UK Highway Code. The SDC is unable to stop in time. Text appears on-screen to inform participants that the SDC could not stop in time and hit the pedestrian who sustained minor injuries.

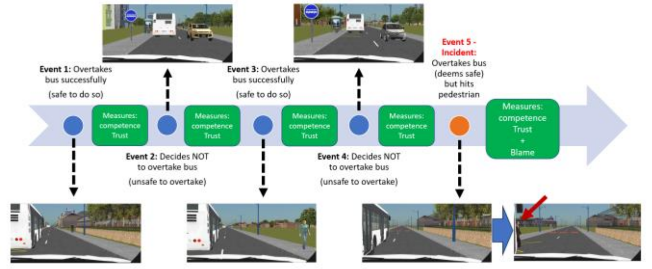
Event 1: Overtakes bus successfully
(safe to do so)
Event 3: Overtakes bus successfully
(safe to do so)
Event 5 - Incident:
Overtakes bus (deems safe) but hits pedestrian
Measures: competence Trust
Measures: competence Trust
Measures: competence Trust
Measures: competence Trust
Measures: competence Trust + Blame
Event 2: Decides NOT to overtake bus
(unsafe to overtake)
Event 4: Decides NOT to overtake bus
(unsafe to overtake)

Figure 1. The New Paradigm

Informational Assistants. These are manipulated through the presence of a Nao robot (*Softbank Robotics*) in half of the conditions and no robot in the others. Dialogue involves regular updates about the SDC intentions (e.g. not looking for opportunities to overtake when unsafe to do so, attempting to overtake when traffic conditions allow) and other aspects of the scenario (e.g. approaching a moving bus). The dialogue conditions differed: informational was 'third person' – e.g. 'the vehicle is', 'this vehicle is'; and conversational 'first person' – e.g., 'we are', 'I am'.

Dependent Measures. Measures were taken immediately after overtake attempts with Visual Analogue Scales (range 0-100). Trust in the system involved a single question (always presented first) and 12 questions from the Trust in Automated Systems Survey [12]. Blame (on the SDC and pedestrian) was measured. A modified version of the Robotic Social Attributes Scale (RoSAS [13]) focused on competence, warmth, and discomfort. A single measure of perceived risk was taken.

III. Initial Results & Discussion

Every successive event resulted in significantly higher trust ($ps \leq .001$) apart from Event 5, where it plummeted. After Event 5, the presence of the humanoid robot informational assistant decreased trust compared to the no robot condition, although currently this difference is not statistically significant ($p = .079$). Other findings will be presented at the workshop.

Counterintuitively, perhaps, having experienced successive successful (without incident) maneuvers might have increased participants expectations about the capabilities of the SDC: especially in the robot present condition. Participants may have expected the robot to have detected the pedestrian, even though the person did not appear in sight until milliseconds before the accident.

References

[1] SAE International, "SAE Levels of Driving Automation Refined for Clarity and International Audience", 3-May-2021. [Online] Available: https://www.sae.org/blog/sae-j3016-update [Accessed: 23-June-2023]

[2] Gov.UK, "Self-driving revolution to boost economy and improve safety" (19-August-2020) [Online] Available: https://www.gov.uk/government/news/self-driving-revolution-to-boost-economy-and-improve-road-safety [Accessed 23-June-2023].

[3] Automated Driving Systems 2.0: A Vision for Safety. Publication DOT HS 812 442, USDOT, US Department of Transportation, (2017).P. H. Kim, K. T. Dirks, and C. D. Cooper, "The repair of trust: A dynamic bilateral perspective and multilevel conceptualization," Acad. Manag. Rev., vol. 34, no. 3, pp. 401–422, Jul. 2009.

[4] Zhang, Q., Wallbridge, C.D., Jones, D.M. & Morgan, P.L (2021). The blame game: double standards apply to autonomous vehicle accidents. AHFE Conference on Human Aspects of Transportation, 25-29 July 2021. Lecture Notes in Networks & Systems Springer, Cham. 308-314.

[5] Liljamo, T., Liimatainen, H., & Pöllänen, S. (2018). Attitudes and concerns on automated vehicles. Transportation Research Part F: Traffic Psychology and Behaviour, 59, 24-44.

[6] Sweet, M.N., Scott, D.M., & Hamiditehrani, S. (2023) Who will adopt private automated vehicles and automated shuttle buses? Testing the roles of past experience and performance expectancy. Transportation Planning and Technology, 46(1) 45-70.accidents. Lecture Notes in Networks and Systems, 270, 308–314.

[7] Marcinkiewicz, V., Wallbridge, C., Zhang, Q. & Morgan, P.L. (2022). Integrating humanoid robots into simulation software generated animations to explore judgments on self-driving car Accidents. Presented at: IEEE RO-MAN 2022, Naples, Italy, 29 Aug - 2 Sep.

[8] Wang, M., Lee, S.C., Sanghavi, H.S., Eskew, M., Zhou, B., & Jeon, M. (2021). In-vehicle intelligent agents in fully autonomous driving: The effects of speech style and embodiment together and separately. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21). Association for Computing Machinery, New York, NY, USA, 247–254.

[9] Wang, M., Lee, S.C., Montavon, G., Qin, J., & Jeon, M. (2022). Conversational Voice Agents are Preferred and Lead to Better Driving Performance in Conditionally Automated Vehicles. In Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '22). Association for Computing Machinery, New York, NY, USA, 86–95.

[10] Onnasch, L., & Hildebrandt, C. L. (2022). Impact of Anthropomorphic Robot Design on Trust and Attention in Industrial Human-Robot Interaction. ACM Transactions on HRI, 11(1).

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

[12] Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA

2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors & Ergonomics 1st ed, 13–30). Springer.

[13] Carpinella, C.M., Wyman, A.B., Perez, M.A., & Stroessner, S.J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and Validation. 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, 254-262).