

Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving

Daniel Omeiza¹, Konrad Kollnig¹, Helena Web¹, Marina Jirotko¹, and Lars Kunze²

Abstract—Autonomous vehicles (AVs) have the potential to change the way we commute, travel, and transport our goods. The deployment of AVs in society, however, requires that people understand, accept, and trust them. Intelligent explanations can help different AV stakeholders to assess AVs’ behaviours, and in turn, increase their confidence and foster trust. In a user study (N = 101), we examined different explanation types (based on investigatory queries) provided by an AV and their effect on people using the trust determinant factors. Our quantitative and qualitative analysis shows that explanations with causal attributions improved task performance and understanding when assessing driving events but did not directly improve perceived trust. This underlines the potential need for additional measures and research to enhance trust in AVs.

I. INTRODUCTION

According to a recent report from the Association for Safe International Road Travel, about 3,000 lives are lost every day in traffic accidents. In efforts to reduce these deaths, advancements in vehicle safety technologies, including speed limits, air bags, crash tests, and seat belts, have contributed remarkably to a reduction in traffic fatalities [1]. Despite these advancements, time spent in traffic continues to be dangerous, with human error being the leading cause for traffic accidents [2].

Autonomous vehicles (AVs) promise to be one of the greatest landmarks in road safety technology. As the reliance on human driver decreases, AVs will reduce road accidents caused by human error, traffic congestion (e.g., those caused by ‘stop and wave’ actions of human drivers), and even the emission of poisonous gases into the atmosphere, while at the same time improving the quality and productivity of the time spent on the road [2], [3]. However, confidence, acceptance and trust in AV will be essential for the deployment of AVs, and these have been challenged by recent events. Prominent examples are the collision of an Uber-owned AV with a pedestrian [4], an automated driving system (ADS) crash caused by rain [5], Tesla’s collision with a truck killing a driver [6], and Google’s ADS failure to correctly estimate the speed of a bus leading to collision [7]. Such incidents negatively affect users’ confidence and trust. Meaningful explanations of AVs’ actions to passengers and other stakeholders (identified in [8]) can play an essential role in the adoption of the AV technology. Moreover, explanation

TABLE I: Explanation types and their investigatory queries. In this work, we study their effect on human perceptions of AVs.

Type	Class	Example Query
Contrastive	Causal	Why Not: why did you not do Y?
Non-Contrastive	Causal	Why: why did you do X?
Counterfactuals	Causal	What If: what would you do if Z?
Informative	Non-Causal	What: what are you doing?

provisions for an AV’s actions can enhance accountability, and in turn, improve trust.

In this paper, we describe a between-group user study carried out to investigate the effect of explanations (with and without causal attributions) provided by an AV to users in a range of driving scenarios. Participants were asked to engage with sequences of images illustrating different driving events with corresponding explanations. We assessed participants’ understanding of the events and their perceptions of AVs using the trust determinants (such as sense of reliability and safety among others). Explanations with causal attributions (causal explanations) are those that explicitly state reasons for an event [9]. We refer to explanations that do otherwise as explanations without causal attribution (non-causal explanations), see Table I. We found that (1) contrastive explanations (responses to *Why Not* queries) are the most effective causal explanations in driving scenarios (non-causal explanations are least effective); and (2) explanations increase users’ understanding and task performance. However, they may not directly improve users’ trust in AVs.

II. RELATED WORK

Several constructs for reasoned explanations grounded in the research literature from relevant fields have been proposed. Wang et al. [10] proposed a conceptual framework detailing how causal filters, that describe how human reasoning processes, inform explanation techniques. These causal filters include contrastive and counterfactual explanations. Mittelstadt et al. [11] argued that the risk of conflicts in communicating explanations, when the explainer (explanation provider) and the explainee (explanation recipient) have different motives, may be mitigated through contrastive, selective, and social explanations. Boris Kment [12] posited that counterfactuals are helpful in explaining, yet the approach to explainability differs with respect to the explainee. Although developers and regulators can benefit from the explanations meant for end-users, they may still require explanations that adapt appropriately to their needs. In the light of explanations targeted towards end-users, a few studies

¹Daniel Omeiza, Konrad Kollnig, Helena Webb, and Marina Jirotko are with the Dept. of Computer Science, University of Oxford. Email: daniel.omeiza@cs.ox.ac.uk.

²Lars Kunze is with Oxford Robotics Institute, Dept. of Engineering, Science, University of Oxford. Email: lars@robots.ox.ac.uk.

investigated explanations in relation to acceptance, willingness to use and trust AVs. Wiegand et al. [13] evaluated what mental models people have of autonomous driving and provided post-hoc explanations of the vehicle’s behaviour based on these models. These authors showed that the display of detected objects and their predicted motion was most important to understand a situation. Post-hoc explanations provided on this premise significantly increased the user’s level of situation awareness. Ha et al. [14] and Koo et al. [15] examined the effect of explanations on peoples’ trust through user studies. Ha et al. [14] examined two explanation types, simple and attributional, as well as perceived risk on trust in AVs in four autonomous driving scenarios with different levels of risk. Their results show that the explanation type significantly affects trust in autonomous vehicles and that, under high levels of perceived risk, attributional explanations lead to the highest trust.

The focus on trust in many existing human-centric research literature on AVs highlights the importance of trust in AVs. If trust is absent, users may refuse to use AVs even when verifiably safe and efficient. This may diminish the wide benefits of AVs for traffic safety.

Our work contributes to existing research by systematically investigating categories of driving scenarios, and providing four different types of explanations (carefully constructed with a schema) in several examples of the scenario categories. We adopted an objective measure (performance on tasks) to evaluate explanation effects. Our work also examines explanations against the different trust determinant factors set up in [16].

III. USER STUDY

Our study methodology included a design that allow participants to learn by engaging, get tested and provide feedback on certain events of an AV. The learning process involved the presentation of a sequence of images of driving scenarios with explanations provided as captions. The testing process followed the same procedure as the learning procedure, but the explanations were replaced by questions about the graphical scenarios.

We investigated 4 different types of explanations (*Why*, *Why Not*, *What If*, and *What*) based on investigatory queries, see Table I. Hence, we setup an online between-group study with 4 groups. We gained approval from the University of Oxford research ethics committee to conduct the study.

A. Participants

Using the Prolific Academic platform¹, we recruited 101 participants. We added filters to ensure participants are 18 or older, resident in the United Kingdom, and fluent in English. 62 participants were female and 39 were male. Their highest educational attainment was: a high school diploma/A-level (29), ongoing undergraduate studies (12), a bachelors degree (48), a post-graduate degree (12). 95 participants possessed at least one form of driving licence, while 6 did not. Asking

participants how many days they drove in a typical week before the COVID-19 pandemic, 16 participants indicated that they drove all 7 days in the week, while 19 of them indicated that they did not drive or would not drive at all in a whole week. Participants took 38 minutes on average to complete the study. Each participant was paid £10 on completion.

B. Survey Structure

1) *Pre-AV Experience*: Each of the 101 participants was randomly assigned to one of four groups: *Why* (N = 27), *Why Not* (N = 24), *What If* (N = 24), and *What* (N = 26). A questionnaire (pre-AV experience questionnaire) was presented to all groups to capture their perception of trust, safety and reliability in AVs before the learning phase. The pre-AV experience questionnaire contained 8 questions with a 5-point Likert scale adapted from a psychometric trust scale recommended by Hoffman et al. [16]. The Hoffman trust scale was chosen after a review of other measurement scales in [17], [18]. The statements tested whether users agree that AVs are rule-abiding, predictable, reliable, safe, efficient, warying, effective, and adoptable by users. The statements were worded in the form: “I currently have confidence in autonomous vehicles and I feel that they obey road rules and can respond appropriately to traffic situations.” Also, participants were asked to provide free responses about what they think of AVs: “What do you think about autonomous vehicles? (e.g. trust, safety, reliability,...)”. Although some authors see trust as a subjective/emotional state, the measurements used here (cf. [16]) focused on the system aspect of trust, that is, how the performance of the AV may affect confidence in the AV.

C. AV-Experience

Each of the four groups was presented with the same sequence of images, illustrating driving scenarios, but with different types of textual explanations (i.e., *Why*, *Why Not*, *What If*, and *What* explanations) as captions for each of the depicted scenario and group. Participants observed the driving events by looking at the image sequences and the corresponding textual explanations explaining the events in the scenarios. We describe the different driving scenarios in the next Section III-C.1.

1) *Driving Scenarios*: In this paper, a scenario represents a set of images. These images illustrate an event with or without explanations. We categorised the driving scenarios into two groups: *goal-oriented actions* and *stimulus-driven actions* [19]. Goal-oriented actions refer to actions that involve the manipulation of the vehicle in navigation tasks such as left turn, right turn, branch and merge. In contrast, when the vehicle is in operation, it can make a stop or deviate decisions due to traffic participants or obstacles. Such decisions are categorised as stimulus-driven actions. For each of the driving action categories, we created *normative events*, *near-misses*, *collisions*, and *emergency events* all occurring in the real-world. These events (especially near-

¹<https://www.prolific.co/>

misses and collisions) are critical and were all accompanied by explanations.

In this study, scenarios were carefully selected to include different AV driving actions (i.e., goal-oriented and stimulus-driven actions) and their corresponding events (i.e. normative, near-miss, collision, and emergency). These different dimensions of actions and events were formed from varieties of left turn and lane merge examples. There was an AV in every scenario, represented as a blue vehicle. There was a total of 24 scenarios used in this stage. Participants were asked to imagine that they were passengers in the AV, and that the explanations were generated by the AV.

2) *Explanation Generation*: After a detailed analysis of driving scenarios, we noticed that the presentation forms of the various causal explanations vary with respect to the driving event under consideration (different types for different uses [20]). To ensure consistency of explanation forms across events, we created an explanation schema for the different event types. Having identified the key elements required to articulate explanations, we carefully designed the schema to appropriately place the elements needed for good intelligibility. For example, the schema for a ‘Why-Not’ explanation for a near-miss event is:

“We [*didn’t/can’t/couldn’t/aren’t*] do [*x*] because [*state the unexpected circumstance*]. [*State the road rule or road sign that apply*].” and the resulting explanation is:

“We can’t move because a vehicle from the side-road unexpectedly moved into the main road obstructing our path.”

D. Post-AV Experience

In this stage, we designed two evaluation measures: the task performance measure and the post-AV questionnaire on the trust factors. The task performance measure was objective while the post-AV questionnaire was a subjective measure.

1) *Task Performance Measure*: The participants were asked to perform some tasks in form of a quiz after interacting with the scenarios and explanations. The tasks comprised 30 questions each in objective form. Each question required the choice of one of four answer options where only one was correct. The tasks also included scenarios that exhibited the different AV driving action classes as well as the corresponding events. The tasks were designed to reflect three forms of questioning styles (which we also refer to as *task categories*) with 10 questions in each category.

- 1) *Prediction*—a single image about a traffic scenario is displayed without an explanation and the participant is asked to predict the next action of the AV.
- 2) *Accountability*—the participant is asked to identify the road participant, who is causing a collision or near miss, in a presented graphical traffic scenario without an explanation.
- 3) *Situation Assessment*—a graphic about a traffic scenario is presented along with four statements that relate to the current scenario. Participants were asked to select one out of the four options that mostly supported the scenario.

2) *Post-AV Experience Questionnaire*: In this stage, the participants were asked to respond to a questionnaire similar to the pre-AV experience questionnaire. However, all the questions in the post-AV experience questionnaire were conditioned on the explanations provided in the AV experience stage. The statements were designed to test whether the participants’ views on AVs being rule-abiding, predictable, reliable, safe, efficient, warying, effective, and adoptable have changed after the study. The statements were worded in the form: “*Based on the explanations provided by the autonomous vehicle in this survey, I have increased confidence in autonomous vehicles and I feel that they obey road rules and can respond appropriately to traffic stimulus*”. Participants were also asked to provide free responses about what they generally think of AVs. The question was “*Based on the explanations provided by the autonomous vehicle in this survey, what do you think about autonomous vehicles? (e.g. trust, safety, reliability,...)*”.

IV. QUANTITATIVE RESULTS

A. Performance in Tasks

To examine the impact of explanations on participants, we set up a set of tasks to test the participants’ understanding of driving events. We used ANOVA and Tukey’s post-hoc paired tests to analyse the performances in the tasks. We assumed that participants’ performances gauge their level of understanding of the AV events. We discovered that the explanation type significantly affected the participants’ understanding of AVs’ events and influenced their performances in the tasks (tasks $F(3, 97) = 8.011, p < 0.001$). Observing the group range scores across driving scenarios, emergency and collision events had the largest range scores, see Figure 1. This suggests that the provision of explanations and the type of explanation is most important in emergency and collision events. The descriptive statistics ($M = 17.8, 20.2, 15.5, 16.1, SD = 4.03, 4.43, 3.12, 2.94$) represent the means and standard deviations for the *Why, Why Not, What If, and What* groups respectively. Participants in the *Why Not* group performed better than those in *What* and *What If* groups, see Figure 1.

B. Trust Factors

We computed the difference in means for each participant’s responses to the eight Likert statements in the pre-AV experience and the post-AV experience questionnaires. We checked if the difference was significant for the participants across the groups. Our results indicate the absence of significant statistical differences, and showed no correlation between the mean differences and the task performance scores ($\rho = 0.037, p = .71$). The participants perspective of the trust factors in AVs mostly declined in the post-AV experience stage. The number of participants, who indicated that they would like to start using AVs for travelling, reduced in the post-AV experience stage, see Figures 4 and 2. We found no correlation between task performance scores and trust difference (i.e. difference between pre-AV experience and post-AV experience) values.

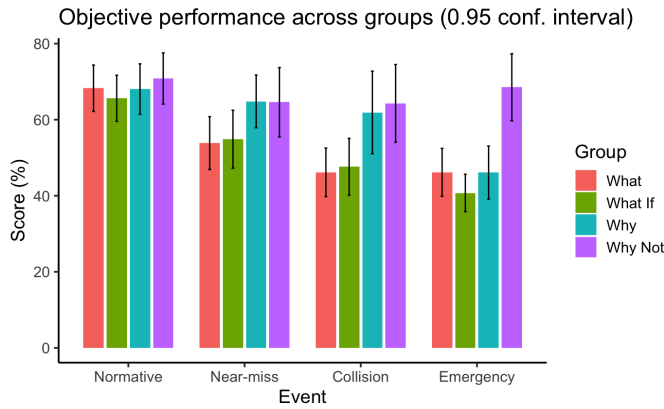


Fig. 1: Task performance in the different *driving events*. With the exception of the near-miss category, participants in the *Why Not* group consistently outperformed other groups. Impact of explanation types was greatest in collision and emergency events.

V. QUALITATIVE RESULTS: THEMES AND REFLECTIONS

A. Pre-AV Experience and Post-AV Experience Evaluation

We performed an inductive thematic analysis on the qualitative (free response) data. Many of the themes occurred across the different explanation groups. In most of the examples provided here, we recorded each participant’s response in the pre-AV experience questionnaire and the post-AV experience questionnaire for easy tracking of changes in comments. Generally, there was an indication of decline in the number of positive comments supporting AVs. We discuss the themes in two broad categories: distrust and trust. See Figure 3 for a frequency plots of trust and distrust comments.

1) *Distrust*: Participants perception of AVs in terms of the trust factors did not improve, in fact, a decline was noticed.

a) *Unwillingness to Give-Up Control*: Although some participants seemed to have increased their understanding of AVs after the AV experience stage, their perception of AVs in terms of the trust factors did not improve. They still prefer to have a full driving control of their vehicles:

“I would be very wary of them. I guess I don’t know enough about them so at the moment don’t feel I would trust them and would prefer to be in control.”–MC (pre-AV experience)

“I still don’t feel confident that they are a reliable and safe way of driving. I would prefer full control of the vehicle.”–MC (post-AV experience).

b) *Too Early to be Trusted*: The nascent nature of AV technologies made some people undermine their current capabilities, thinking that they are still too early to be trusted:

“I think that the technology is not sufficiently advanced to make them safe enough to use.”–LT (pre-AV experience).

“I’d not use them. It needs many years of other people using them to convince me that they are safe.”–LT (post-AV experience).

c) *Worry about Reliability and Robustness*: Although, there were a couple of assurance from participants on

AVs’ compliance to road rules in the post-AV experience, there were worries about their reliability and response to unpredictable situations. Participants were unsure about the efficiency of the take-over process in such unpredictable situations:

“I have four big concerns: 1) How reliable is the current technology; Could they spot and avoid a child dressed in dark clothes who runs out into the road during a thunderstorm as effectively as a human? [...] 3) Reliability. What if the vehicle malfunctions; A big malfunction is easy to spot, but what about something subtle? There might be no obvious problem until there’s an accident [...]”–NG (pre-AV experience).

“I feel that they are designed to follow precise rules and in theory should be safe. However, I still feel uncomfortable about how they might respond in unpredictable situations which haven’t been programmed in (e.g. another vehicle driving [erratically]) or whether they could spot a potential hazard (e.g. a girl playing with a ball who might potentially run out into the road).”–NG (post-AV experience).

d) *Track Records of Accidents*: Apparently, previous records of accidents in the news, and the few collision examples (due to offence by other road participants) shown to the participants in the AV-Experience stage of the study increased participants’ doubts: “I would not use one until they are firmly established into society and have a proven track record of safety.”–DA (post-AV experience).

“From the explanations the AV mostly complied with the road rules but on the odd occasion there was still [a] collision. I would expect an AV to [avoid] collisions more often than crashing into the other vehicle”–TJ (post-AV experience).

e) *Worry about Environment, Security, and Prioritisation*: Participants thought that the dynamic nature of driving environments, prioritisation of road participants (not just emergency vehicles as shown in the scenarios) and security of the AVs were crucial and might have affected their trust factor rating in the study:

“[are] safe than [I] originally thought, but I am concerned that there will be roads such [as] very narrow ones where normal rules [won’t] apply like single car only plus if there was a person in the road.”–DR (post-AV experience).

“No change to my initial thoughts. Still worried that AVs could be compromised in relation to security”–HC (post-AV experience).

f) *Ethical Concerns*: Some participants raised ethical concerns about the design process of AVs:

“Ethics. Who programs the vehicle to make choices [...] ?”–NG (post-AV experience).

“It has also not gone into any depth about moral decisions”–BJ (post-AV experience).

2) *Trust*: Some of the participants that received explanations with causal attributions indicated that the explanation was helpful in explaining the reasons behind the driving decisions of the AV and therefore somewhat trust them. In fact, some participants in the *Why Not* group had a change of view in favour of trust.

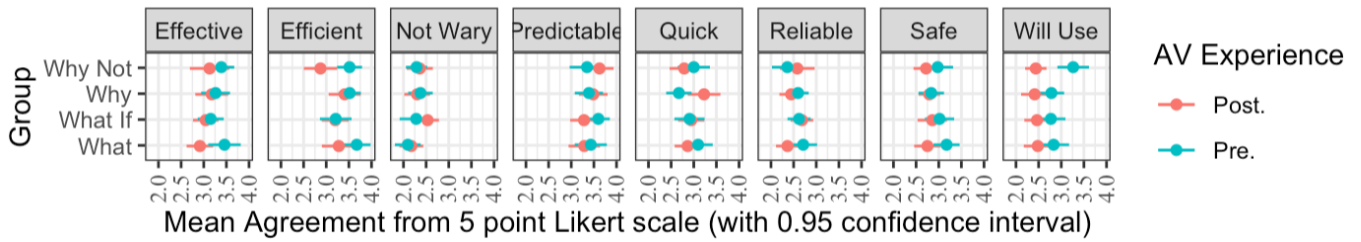


Fig. 2: Mean values for each of the trust factors represented in the pre-AV experience (represented with the green plot) and post-AV experience (represented with the pink plot) questionnaires. The box captions indicate the different trust factors. The pre-AV experience mean values were higher but there was no significant difference between the pre-AV and post-AV experience mean ratings across groups.

a) *Explanations Enhanced Trust*: “I don’t understand how they can react quickly enough in an emergency situation to be safe. I am very wary!”–KP (pre-AV experience).

“The explanations were very clear, and I could see the reasoning behind the driving decisions which were made, which has reassured me somewhat. Maybe AVs are safer than I think.”–KP (post-AV experience).

b) *AVs are more Efficient than Humans*: Some of the participants made a comparison with human drivers. They suggested that AVs obey rules better than human drivers and would mostly be efficient when there are few human drivers on the roads:

“More safe than human drivers but can’t respond to every situation yet”–DK (pre-AV experience)

“I think they understand the rules of the road better than some human drivers”–DK (post-AV experience).

c) *Improved Productivity and Appeal for Publicity*: Some also thought AVs could be well-suited for long journeys, and that designers and manufacturers needed to promote this benefit better for a widespread adoption:

“I think they would be an excellent way for me to make more out of my day. 2 hours commuting time where I may be able to do other things than concentrate on the road”–CA (pre-AV experience). The participant remained positive and stresses the need for more AVs. “I think that they can be trusted, but a lot more data needs to be collected before we get there. Also I think it will be safer for all cars to be autonomous rather than some [autonomous] and some normal drivers”–CA (post-AV experience).

Some suggested that AVs and their benefits should be promoted for wider adoption and trust: “I think its going to be a big part of the future. But to get [there] we need to project and appeal to mass trust from society by ensuring safety and reliability”–FR (pre-AV experience). “I think AV manufacturers and companies should promote [AVs] more to the public to get better widespread perceptions”–FR (post-AV experience).

VI. DISCUSSION

Drawing on the findings, situational awareness is positively impacted by concise and clear causal explanations, in particular by the contrastive (or *Why Not*) type. This means that explanations in AVs should be constructed with reference to relevant foils such as other road participants

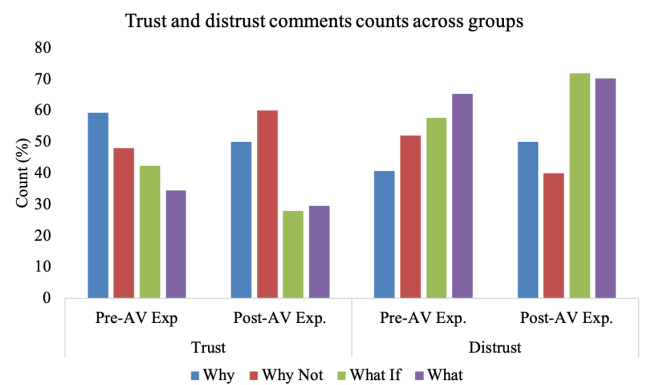


Fig. 3: Frequency of negative (distrust) and positive (trust) comments about trust in AVs. The *y*-axis indicates the frequency in percentage, while the *x*-axis indicates the pre-AV and post-AV experiences along with trust and distrust comments. Only the *Why Not* group had increased positive comments in the post-AV experience questionnaire.

(e.g., pedestrians, cyclists, other vehicles), traffic signs, and routes.

A. Perceptions of AVs

Explanations are thought to have a relationship with trust [21], [22]; however, they might not directly create or improve trust as observed in this study. In the quantitative analysis of the trust factors, the mean rating values generally reduced. The qualitative data suggests some background why this was the case. From the qualitative data, many participants, who made negative comments about AVs in the pre-AV experience stage, also made negative comments in the post-AV experience stage but with reasons that indicated their enhanced knowledge and experience of AVs. Interestingly, participants in the *Why Not* group made increased positive comments about AVs, see Figure 3. Participants in the *Why Not* group, who indicated neutral on the Likert scale in the pre-AV questionnaire, must have made comments that favoured trust in the free response section. With the aim to have representative scenarios, we added a few collision examples in the learning and task stages. In all the collision cases, the non-autonomous vehicles violated the traffic rules in a constrained environment so that the AV did not have the chance to adjust immediately. This is a reason for the decrease in trust supporting comments as mentioned by a

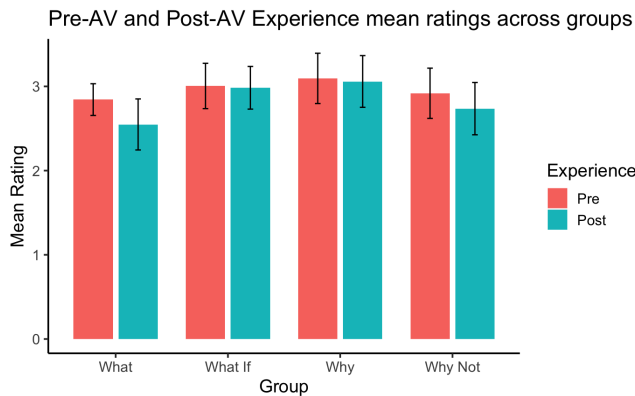


Fig. 4: Trust factors evaluated through the pre-AV (pink) and the post-AV experience questionnaire (green) with 5-point Likert scales. The y -axis indicates the mean response values (high is strongly agree, low is strongly disagree) for all the trust factors combined while the x -axis indicates the respective groups. Non-causal explanation (*what group*) had highest decline.

few participants among other reasons (e.g., lack of real-world first-hand experience and previous accident cases). The participants in the *Why Not* group had an increased number of trust supporting comments. Further, the trust factors measurement adopted was primarily linked to the performance of the AV. In any case, the role of explanation still remains crucial as participants had increased understanding of AVs through the explanations. In addition, collisions would hardly occur when all or most of the vehicles on the road are autonomous.

B. Regulations and Standards

Regulators have a role to play in ensuring trustworthy autonomous vehicles. Some of the concerns raised in the study touch on effective testing by relevant stakeholders to ensure safety. Increased efforts towards strengthening existing related standards and setting new ones to help explainability and address the ethical concerns around AVs' adoption (as raised by some participants) are needed. This will enhance accountability and trust in AVs.

VII. CONCLUSION

We described a user study to investigate the effect of explanations on humans in different autonomous driving conditions. Our findings show that providing explanations with causal attributions, and in particular, contrastive (or *Why Not*) explanations, can improve the understanding and situation awareness of AV passengers. Our results showed that the explanation type was more significant in emergency and collision events. However, perceived trust of the subjects in AVs declined after engagement with the AV. We detailed the reasons why people might not currently trust AVs. In future work, we will investigate whether our results can be confirmed when using a high-fidelity prototype for a more immersive AV experience. Overall, we believe that this study provides valuable insights on the effects of explanations on human perceptions of autonomous vehicles.

ACKNOWLEDGMENT

This work was supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) through project RoboTIPS: Developing Responsible Robots for the Digital Economy, grant reference EP/S005099/1. It was also supported by the Assuring Autonomy International Programme (Demonstrator project: Sense-Assess-eXplain (SAX)), a partnership between Lloyd's Register Foundation and the University of York.

REFERENCES

- [1] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human-Computer Interaction*, vol. 31, no. 10, pp. 692–702, 2015.
- [2] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, 2018.
- [3] P. Goldin, "Advantages of autonomous vehicles," 10.
- [4] N. A. Stanton, P. M. Salmon, G. H. Walker, and M. Stanton, "Models and methods for collision analysis: a comparison study based on the uber collision with a pedestrian," *Safety Science*, vol. 120, pp. 117–128, 2019.
- [5] D. Lavrinc, "This is how bad self-driving cars suck in rain," "2014 (accessed July 24, 2020)".
- [6] M. McFarland, "Who's responsible when an autonomous car crashes?," "2016 (accessed July 24, 2020)".
- [7] A. Davies, "Google's self-driving car caused its first crash," "2016 (accessed 24 July, 2020).
- [8] D. Omeiza, H. Webb, M. Jirotko, and L. Kunze, "Explanations in autonomous driving: A survey," *arXiv preprint arXiv:2103.05154*, 2021.
- [9] H. H. Kelley, "The processes of causal attribution.," *American psychologist*, vol. 28, no. 2, p. 107, 1973.
- [10] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15, 2019.
- [11] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.
- [12] B. Kment, "Counterfactuals and explanation," *Mind*, vol. 115, no. 458, pp. 261–310, 2006.
- [13] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, and H. Hussmann, "I drive-you trust: Explaining driving behavior of autonomous cars," in *Extended abstracts of the 2019 chi conference on human factors in computing systems*, pp. 1–6, 2019.
- [14] T. Ha, S. Kim, D. Seo, and S. Lee, "Effects of explanation types and perceived risk on trust in autonomous vehicles," *Transportation research part F: traffic psychology and behaviour*, vol. 73, pp. 271–280, 2020.
- [15] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *International Journal on Interactive Design and Manufacturing (IJDeM)*, vol. 9, no. 4, pp. 269–275, 2015.
- [16] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.
- [17] 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.
- [18] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, vol. 53, pp. 6–8, Citeseer, 2000.
- [19] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7699–7707, 2018.
- [20] Y. Zhou and D. Danks, "Different" intelligibility" for different folks," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 194–199, 2020.

- [21] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proceedings of the 11th international conference on Intelligent user interfaces*, pp. 93–100, 2006.
- [22] W. Pieters, "Explanation and trust: what to tell the user in security and ai?," *Ethics and information technology*, vol. 13, no. 1, pp. 53–64, 2011.