Towards Accountability: Providing Intelligible Explanations in Autonomous Driving

Daniel Omeiza[1], Helena Web[1], Marina Jirotka[1], and Lars Kunze[2]

Abstract— The safe deployment of autonomous vehicles (AVs) in real world scenarios requires that AVs are accountable. One way of ensuring accountability is through the provision of explanations for what the vehicles have 'seen', done and might do in a given scenario. Intelligible explanations can help developers and regulators to assess AVs' behaviour, and in turn, uphold accountability. In this paper, we propose an interpretable (tree-based) and user-centric approach for explaining autonomous driving behaviours. In a user study, we examined different explanation types instigated by investigatory queries. We conducted an experiment to identify scenarios that require explanations and the corresponding appropriate explanation types for such scenarios. Our findings show that an explanation type matters mostly in emergency and collision driving conditions. Also, providing intelligible explanations (especially contrastive types) with causal attributions can improve accountability in autonomous driving. The proposed interpretable approach can help realise such intelligible explanations with causal attributions.

I. INTRODUCTION

The increasing growth rate in the automotive industry is precipitated by the accrued research knowledge in vehicle dynamics, the emergence of deep learning algorithms, the development of new and enhanced sensing devices (as described in [1]), and possible market potential [2]. Despite the technological advancements, the successful deployment of AVs in the real world may greatly depend on users' acceptance and confidence. Due to reports on AV accident cases [3], [4], public skepticism in the acceptance of AVs in society still seems to exist. Effective means to building public confidence in AVs are therefore necessary.

A means of building confidence and increasing public acceptance is through the provision of explanations. AVs make high-stake decisions that can significantly affect humans. Hence, they should *intelligibly* explain and justify their decisions sufficiently to uphold *accountability*. Most of the existing explanation techniques mainly focus on explaining data-driven models (e.g., machine learning models) with less attention on complex goal-based systems such as AVs. Moreover, the explanations they offer also suffer from low intelligibility [5], [6]. This makes them mainly beneficial to the experts and not readily utilisable by lay users. In addition, only a few human-centric studies on explanations have been conducted in the autonomous driving context.

[1]Daniel Omeiza, Helena Webb, and Marina Jirotka are with the Dept. of Computer Science, University of Oxford. Email: daniel.omeiza@cs.ox.ac.uk.
[2]Lars Kunze is with Oxford Robotics Institute, Dept. of Engineering, Science, University of Oxford. Email: lars@robots.ox.ac.uk.
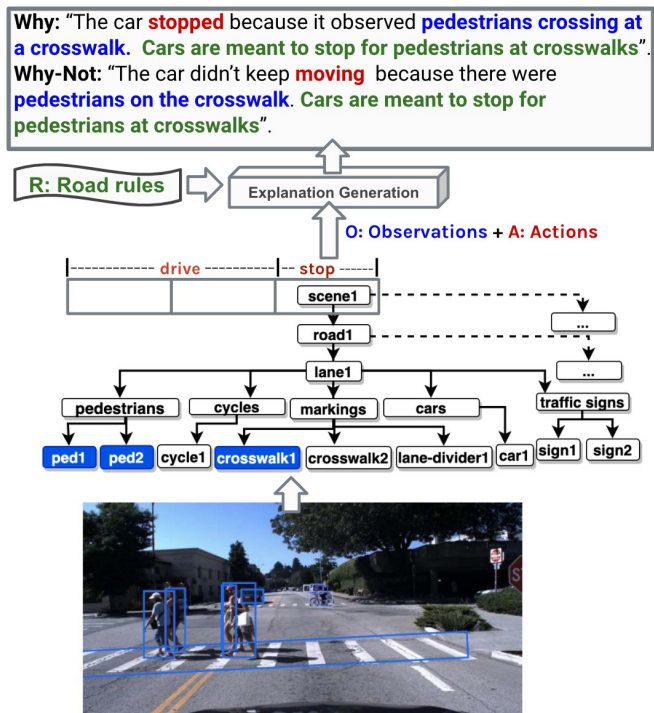
Fig. 1: Interpretable Representation. Different types of explanations (e.g. *Why* and *Why Not*) generated from the underlying representations of actions (A), observations (O), and road rules (R). The observations and actions are obtained from scene graphs representing various frames from a driving scene video. The actions (A), observations (O), and road rules (R) are used to build a tree in the 'Explanation Generation' phase. We focused on the tree representation in the 'Explanation Generation' phase and the impact of the explanations on humans. Using a user study, we evaluated the impact of the generated explanations in a range of driving scenarios and assessed them against intelligibility and accountability goals.

Using *intelligibility* and *accountability* goals, this research proposes an interpretable and user-centric method to explanation provision in autonomous driving. It builds on risk object identification technique in driving scene [7] and traffic objects representation using scene graphs [8] (see Figure 1). The observations recorded in the scene graphs, the AV's actions, and the road rules are combined to generate explanations. In this paper, we only focus on how this combination can generate different explanations (i.e., the proposed tree-based method), and the type of driving conditions where the explanations are mostly appropriate (i.e. through a user study).

Our main contributions are:
- We propose an interpretable tree-based representation for generating different types of explanations based on

observations, *actions*, and *road rules*.

- We present a user study that evaluates different types of explanations (generated by following the proposed tree-based approach) in a range of driving scenarios. In particular, we show the importance of causal explanations in safety-critical scenarios.
- In contrast to existing work on explanations in AVs, we apply a triangulation design method (i.e., both quantitative and qualitative methods) to evaluate explanations for autonomous driving against intelligibility and accountability goals.

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.

II. BACKGROUND

Explanations have been studied in various contexts to determine their impact on people [10], [11]. As the effectiveness of explanations differs with domains and their context [12], investigating explanations in the context of autonomous driving is key.

In the recent work, deep learning model have been trained on video driving data with textual explanations as annotations [13], [14]. Only a few works employed user studies as a means to examine the impact of explanations on stakeholders in the driving context [15]. Ha et al. [16] and Koo et al. [17] examined people's perception of trust in autonomous vehicles by conditioning the participants to explanations presumed to have been provided by an AV.

We provide a background for intelligibility and accountability as we look at explanations with intelligibility and accountability goals in mind.

A. Intelligibility

Article 12 of the GDPR [18] demands that information be provided to data subjects in an *intelligible* construct. The term intelligibility is used to describe how easy an explanation could be understood or comprehended [11]. While some existing artificial intelligent systems equipped with explanation mechanisms could be beneficial to experts, a thorough investigation of their explanation properties show no indication of *intelligibility* when lay users are involved [19]. Further, many explanation algorithm design processes are not informed by users' needs. For example, Chakraborti et al. [20] proposed an algorithm for explaining the plans of an autonomous robot. The explanations generated by this algorithm were not communicated in natural language. Therefore, they are not easy to comprehend by lay users. This is a serious concern, especially for AVs and social robots where the party who mainly requires explanations is usually not an AI expert. In fact, visual explanations in the form of saliency maps may not pass clear or correct messages [21] to an AV auxiliary driver who needs to immediately act where an intervention is required. Intelligible explanations are therefore critical as it is seen as one way of ensuring accountability in autonomous driving.

TABLE I: Explanation types and their investigatory queries as used in this study.

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?

B. Accountability

Regulators and auditors (e.g., the GDPR) are seeking ways to ensure *accountability* in algorithmic systems. One way they are doing this is by setting authorities to audit algorithmic systems in order to ensure their compliance to the guidelines. Auditing can be challenging in blackbox systems without explanations [22]. Moreover, accident investigation in AI systems, as described in [23], requires meaningful transparency in autonomous systems so as to enable easy investigation by different stakeholders. The high complexity of the explanation techniques (e.g., training a deep learning model with scenes and corresponding explanations) for AVs which have been applied in previous works makes accountability difficult to achieve.

In the other hands, explanations could be overwhelming or redundant. In the next section, we discuss redundancy in explanation.

C. Redundancy

The existing explanation techniques [13], [14] applied in AVs so far focus on explaining only a specific aspect of an AV, which are usually one or two actions (e.g., left turn and forward movement). Often, the aspects focused on are not very critical and explanations might not be crucial for such cases. Explanations for irrelevant scenarios could, in fact, be disturbing to users. One contributing factor to this problem is the widespread use of non user-centric design approaches for explanation systems.

As a step towards more user-centric explanation design, we first describe a transparent and interpretable approach to generating intelligible explanations from *observations*, *road rules*, and *actions* using a tree-based data structure. Further, we describe a user study we have conducted to examine different explanations (see Table I) in different driving conditions using the intelligibility and accountability objectives.

III. EXPLANATION GENERATION

After a careful analysis of different driving scenarios, we identified three variables required to provide an intelligible explanation. These include a set of road rules (R), observations (O), and actions (A). Conditions such as traffic offence, justification for an action, unexpected circumstance, and an AV's reaction to other agents' offence were derived from these three variables: R, O, and A. Observations are objects, agents or road signs in the environment detected by a detection and tracking model. Road rules are part of the domain knowledge of the agent. Rule selection in each
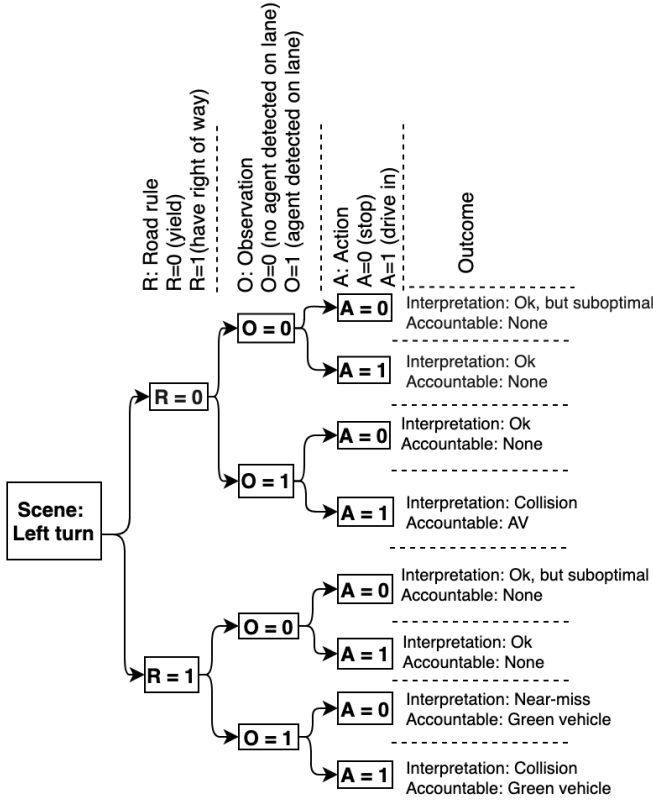
Fig. 2: Underlying tree-based representation used in the explanation generation phase depicted in Figure 1. The tree is constructed with key variables: road rules (R), observations (O), and actions (A). Different types of explanations are generated through different traversals of the tree. We manually interpreted the outcomes indicating accountability (especially in collision incidents) for each path in the tree representation one of the left turn scenarios used in the user study (see Figure 3).

situation is informed by observations. Actions represent the decisions taken based on observations and rules.

Explanation X is represented with a variable size tuples ($X = \langle...\rangle$). For example, $X = \langle A, O, R \rangle$ will yield an explanation of the form:

> The car *[describe action]* because *[describe observed circumstance]* and *[reference the relevant road rules]*.

In each driving scenario, a tree is created from the observations and actions represented in a scene graph, and the road rules (see Figure 1). The tree is traversed to collect the values for each member of the tuple needed to form X. Figure 2 shows an example of a constructed tree using R, O and A for the left turn scenario shown in Figure 3.

Formally, let T represents a tree such that $T = \langle N, E \rangle$. N is a set of nodes and E is a set of edges connecting two nodes. We define a node $n \in N$ in the tree as a tuple $n = (u, S)$ where: $u \in \mathbb{N}$ is a unique numerical identifier for a node in T. S is a list of statements based on $R, O,$ or A (e.g. $A = 1$). The root node is the unique node with no parent, and a leaf is a node with no child. The level l_n of a node n is the number of edges from the root to that node.

We explain how four different explanation types (*Why*,

Why Not, *What If*, and *What*) can be generated from T.

To construct a *Why* explanation, we traverse T by starting from the root node (say $n_r \in N$) to a leaf node (say $n_l \in N$). We return the set of unique statements S_w which satisfy the decision trace of the input instance. *Why* explanation is then created using the information in S_w. Each $s \in S_w$ is then represented with linguistic terms that describe its meaning in the driving domain. The 'Why' explanation X_w is now a concatenation of the linguistic terms for all the $s \in S_w$.

For *Why Not* explanations, traverse T to generate S_w for the *why-trace*. Find and note the lowest common ancestor n_a of leaf node n_l and the foil n_l' (i.e., the closest alternative output). n_a is the node from which the path p_w from the root n_r to n_l and the path p_{wn} from the root n_r to n_l' first differ. Thus, the statement s at n_a is crucial for explaining why n_l' was not obtained. *Why Not* explanation X_{wn} is constructed using the linguistic representation of s.

What If explanations are also referred to as counterfactual explanations. Counterfactual explanations are meant to contain information about the minimum change required to obtain the closest alternative outcome or foil. To construct a *What If* explanation, find the closest foil n_l' to the current leaf node n_l. Obtain the lowest common ancestor n_a of n_l and n_l' in T. The statement s at n_a is negated (e.g., $\neg O = 1$) and added to the set of statements (say S_f) resulting from $p_w \setminus p_{wn}$ (where '$\setminus$' represents set complement). Each $s \in S_f$ is then represented with linguistic terms that describe its meaning in the driving domain. The *What If* explanation X_f is a concatenation of the linguistic terms for all the $s \in S_f$.

To generate *What* explanation, use the leaf node n_l which represents an action statement that corresponds to the scenario being considered.

In contrast to existing works which employed deep learning approaches (e.g. [13], [14], [24], [25]) our proposed interpretable method combines road rules, observations, and actions to realise intelligible explanations for AVs. The tree representation can be generalised with respect to the complexity of the propositions within nodes and the complexity of the scene (tree can have variable depth and branching factor). To rectify accountability errors arising from the misidentification of objects by the perception system, an ethical blackbox [26] (i.e., a transparent and accurate event data recorder) can be leveraged to validate explanations in critical incidents, such as, collision.

We applied the described tree-based approach to generate explanations for the scenarios in our user study.

IV. USER STUDY

We examined the methodological aspects of related work (such as [11], [10]) and adapted a combination of them. Because highly automated vehicles are not prevalent in many communities, only a handful of people have been directly affected by their decisions. Hence, our study methodology included a design for participants to learn by engaging, and get tested on certain events of an AV. The learning process involved the presentation of a sequence of graphical images of driving scenarios with explanations provided as captions.

We introduced new road signs in the scenarios in an attempt to place all participants on a seemingly level ground. 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 sought and gained approval from the University of Oxford research ethics committee to conduct the study.

A. Participants

We recruited 101 participants via the Prolific Academic platform and applied filters to include only individuals over the age of 18, resident in the United Kingdom, and fluent in English language. 39 of the participants were male and 62 were female.

Their educational experiences ranged from: high school diploma/A-level (29), enrolled for bachelor's (12), bachelors degree (48), to post-graduate degrees (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 lock-down, 16 participants indicated that they drove all 7 days in the week before the lock-down, while 19 of them indicated that they didn't drive or would not drive at all in a week. Overall, participants took 38 minutes on average to complete the study, and each participant was paid £10 on completion.

The study was structured in two phases.

B. Phase 1

In this phase, the 101 participants were randomly assigned to 1 of 4 groups: *Why* (N = 27), *Why Not* (N = 24), *What If* (N = 24), and *What* (N = 26). Each group was presented with the same sequence of scene 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 (image captions) which explained the events in the scenario.

1) Driving Scenarios: In this paper, a scenario basically represents a set of images that illustrate an event with or without explanations as captions. In the research literature, driving scenarios can be broadly categorised into two groups. In particular Ramanishka et al. [27] categorised AV driving actions into *goal-oriented actions* and *stimulus-driven actions*. 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, while the vehicle is in operation, it can make a stop or deviate decision due to traffic participants or obstacles. Stop and deviate are categorised as stimulus-driven actions. For each of the driving action categories, we created normative events, near-miss events, collision events, and emergency events, all obtainable in the real-world.
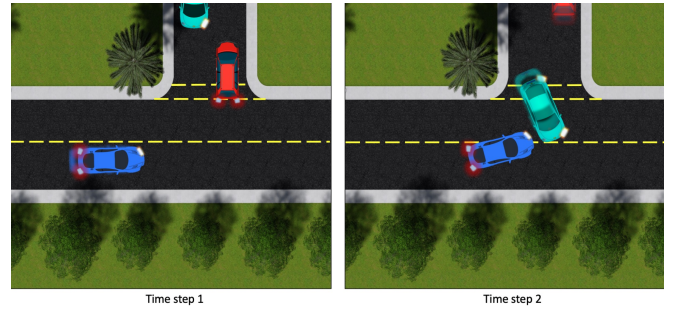

Time step 1   Time step 2

Fig. 3: This scenario from our user study depicts a *near miss in a left-turn* situation involving an autonomous vehicle (AV) (blue) and another traffic agent (green). The green vehicle failed to yield, but the AV adjusted to avoid a collision. Parameters for the scenario depicted are: Observations (O) = {0:no agent detected on lane, 1:agent detected on lane}. Road rules (R) = {0:yield, 1:have right of way}, Action (A) = {0:stop, 1:drive in}. See Figure 2 for the basic tree representation. The following two examples of explanations (generated using the tree representation) were given by the AV to different user groups for the depicted scenario: **Non-Causal Explanation (*What group*):** 'We are stopping.' **Causal Explanation (*Why Not group*):** 'We can't continue because a vehicle from the side-road unexpectedly moved into the main road obstructing our path. The default rule requires that vehicles on the side-road yield to vehicles on the main road.'

1) *Normative* events occur when all road participants including the AV obey the traffic rules.
2) *Near-miss* events occur when a participant violates the traffic or road rules, and the AV has to adjust to avoid a collision.
3) *Collision* events occur when two or more vehicles (including the AV) crash into each other. This occurs when one participant suddenly violates traffic rules and the AV failed to adjust accordingly to avoid an accident.
4) *Emergency* events occur in situations where there is an emergency vehicle such as an ambulance, fire fighters' van, or police van. These emergency vehicles have right of way in all situations, and some of their actions permissively violate default road rules. The AV and other traffic participants are expected to yield in virtually all cases.

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, and it was always a blue coloured vehicle. The total number of scenarios used in this stage was 24. 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 careful analysis of driving scenarios, we noticed that the presentation forms of the various causal explanations vary with respect to the driving event under consideration [12]. To ensure consistency of explanation forms across events, we created an explanation schema for the different event types. We designed the schema to appropriately place the elements needed for good intelligi-

bility. We manually populated the explanation schema using the tree-based method described in Section III.

C. Phase 2

Phase 2 was an evaluation phase. We designed three performance evaluation measures: the task performance measure, the driving rules agreement measure, and the goodness of explanation measure. Some of the measures were objective while others were subjective.

1) Objective Measure: The Quiz Performance: After the interactions with the scenarios and the explanations, the participants were asked to perform some tasks (in the form of a quiz) on similar driving events. The task comprised 30 questions. The questions were in objective form and required the choice of an answer out of 4 options of which only one is correct. The tasks also included scenarios that exhibited the different AV driving action categories 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 caused 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 of the four options that mostly supported the scenario.

2) Objective Measure: Driving Rules Agreement: We stated important road rules that were applied in the learning stage and asked participants to rate their agreements with the rules on a 5-point Likert scale. We assumed that participants with good performance in the tasks would strongly agree with all the statements as the statements were all correct.

3) Subjective Measure: Goodness of Explanation: Participants were provided with 7 statements to elicit the basic properties of good explanations as discussed in [28], [29]; hence, the term 'goodness of explanation'. The participants were asked to rate their agreement with the statements on a 5-point Likert scale.

The goodness of explanation construct employed was founded on those developed in the evaluation metric for explainable AI research summarised by [29] and was adapted to fit our use case. Participants were also asked to provide free responses about what they did not like about the explanations, what they liked, and what they expected of a good explanation.

We carefully formulated hypotheses and expect to validate them using the results from this study.

D. Hypotheses

We hypothesise that different forms of explanations would influence (to varying degrees) the end-users' understanding of AV events.

Our hypotheses for each of the explanation types is hereby detailed:

H1—Intelligibility: Contrastive explanations are preferred by humans because humans generally expect a contrastive response when they ask questions [28] under normal circumstances. We therefore expect that:

> *Why Not* explanations will generally yield the best user understanding over *Why, What If,* and *What* explanations.

H2—Accountability: We also expect that:

> The participants in the *Why Not* group will produce the best performance in the accountability tasks.

We measured the levels of understanding through tasks performance (in the form of a quiz) and a questionnaire on road rules that tests participants' comprehension of the AV events and road rules.

V. RESULTS

A. Quantitative Result

In this section, we present quantitative results relevant to our hypothesis.

1) Test of Hypothesis H1—Intelligibility: We used ANOVA and Tukey's post-hoc paired tests to analyse the performances in the quiz. We assumed that participants' performances gauge their level of understanding of the AV's events. We discovered that explanation type significantly affected the participants' understanding of the AV's events as reflected in the quiz performances (quiz $F(3, 97) = 8.011$, $p < 0.001$). Observing the group range scores across driving scenarios, emergency and collision events had the largest range scores (See Figure 4). Hence, the provision of explanations and the type of explanation is mostly important in emergency and collision events. The descriptive statistic (M = 17.8, 20.2, 15.5, 16.1, SD = 4.03, 4.43, 3.12, 2.94) represent the means and standard deviation 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. Hypothesis **H1** was therefore not rejected. See Figure 4

2) Test of Hypothesis H2—Accountability: We analysed scores based on the task categories (prediction, accountability, and situation assessment). We used ANOVA test to determine the types of explanations appropriate for each of the three categories. We discovered that in the *accountability tasks*, participants in the *Why Not* group had the best performance with a significant difference to *What* group ($p = .01$), *What If* group ($p < .001$), and *Why* group ($p = .007$). The result supported hypothesis **H2**, we therefore did not reject hypothesis **H2**.

B. Goodness of Explanation

We repeated the procedure for the goodness of explanation responses and observed a significant difference in explanation goodness mean ratings across groups ($F(3, 97) = 10.0$, $p < .001$). Means and standard deviations of the goodness of explanation ratings were: (M = 3.83,
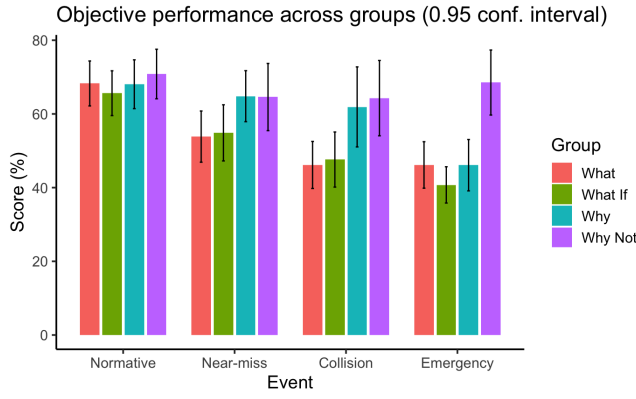
Fig. 4: Quiz task performance in the different *driving events*. With the exception of the near-miss category, participants in the *Why Not* group consistently out-performed the participants in the other groups. Impacts of explanation types was greatest in the collision and emergency events.
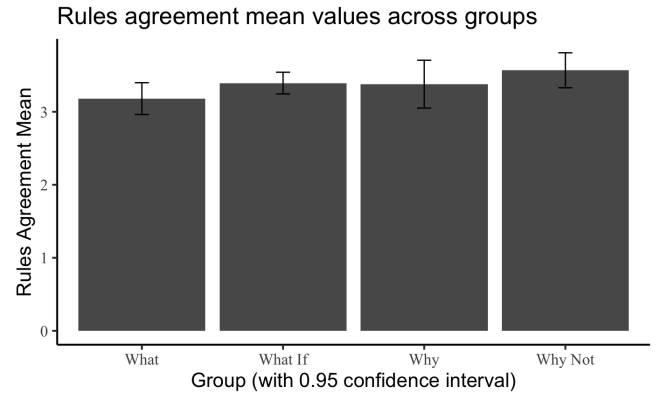

Fig. 5: Rule agreement performance from the rule agreement questionnaire. Participants in the *Why Not* group best understood the rules as they mostly agreed with the road rules.

3.34, 3.25, 2.83, SD = 0.35, 0.65, 0.77, 0.88). The highest mean rating was from the participants in the *Why* group. No correlation was observed between the explanation goodness mean ratings and the quiz scores (rho = 0.19, p = .051).

1) Driving Rules Agreement: Using Pearson's correlation coefficient, we checked the correlation between the mean road-rules agreement rating and quiz scores. The mean agreement values and standard deviations were (M = 3.38, 3.57, 3.39, 3.18, SD = 0.83, 0.57, 0.35, 0.54) for *Why, Why Not, What If,* and *What* respectively (See Figure 5). We found that there was a weak positive correlation between the two variables (rho = 0.34, p < .001). Our result indicated that *Why Not* group had better understanding of the road rules.

C. Qualitative Results: Themes and Reflections

In addition to the Likert scale statements on the goodness of explanation questionnaire, participants were asked to provide free comments. The open-ended questions were:

- 'What are some of the things you like about the textual explanations?'
- 'What are some of the things you do not like about the textual explanations?'
- 'What are the other elements you would like an explanation to have?'

We performed an inductive thematic analysis on the comments from each of the 4 groups. The themes indicated that participants generally prefer short explanations with sufficient information; and that explanation presentation mode matters. Some participants suggested that explanations should be provided as bullet points, and that the AV should provide more details on road signs when explaining events. See Table II for a comprehensive result.

D. Conclusion

Drawing on the findings, *intelligibility* is mostly positively impacted by concise and clear contrastive (*Why Not*) explanations. This means that explanations in AVs should be constructed with reference to relevant foils such as *observed*

road participants (e.g. pedestrians, cyclists, other vehicles), *road rules*, and *actions* especially in emergency and collision events. They should also be sufficiently clear for lay users to comprehend.

In addition, *Why Not* explanations were the most effective in the *accountability* tasks. We conclude that explanations with causal attributions, especially the contrastive types, relatively increase human understanding of AV behaviours and are helpful in upholding accountability. The conceptual tree-based approach proposed in this paper can be used to generate causal and non-causal explanations as explained in this paper.

In summary, we have proposed a conceptual tree-based approach for generating explanations with and without causal attributions. We also described an experiment to investigate explanations in different autonomous driving conditions. Our findings show that providing explanations with causal attributions, and in particular, contrastive (or *Why Not*) explanations, can improve intelligibility and accountability in AVs. Further, the results indicated that an explanation's type is more significant in the emergency and collision events. In future work, we shall introduce probability into the tree and provide formal and experimental proofs of the proposed method. We will also investigate whether our results can be confirmed using a high-fidelity prototype for a more immersive AV experience since some of the events (e.g. collision) cannot be controlled in a real-world setting.

TABLE II: Common themes per group from the participants' comments on the explanations provided.

	Limitation	Strength	Suggestion
Why	- Information overload - Ineffective communication of speed & priority	- Proved that the AV takes the errors of other participants into account - Easy to visualise and imagine - Informative and explained occurrences well	- Explanation of traffic signs - Use of videos - Use of bullet points
Why Not	- Information overload - Situational report and not mechanistic - Road signs unexplained - Not enough clarity	- Simple and easy to follow - Detailed - Shows the 'mechanics' of how things work	- Use of bullet points - Prediction of behaviours - Road signs labelling - Improve clarity on complex scenarios
What If	- Limited information - Too open ended - Difficult to understand	- Visual aids - Explained errors - Travel directions and vehicle gesture representation - Enlightening	- How fast and calculated evasive action would be taken by AVs, when required. - Provided only when necessary - Speed indication and road signs labelling
What	- Not detailed enough - No reasons provided - Hard to figure out road signs - Too short	- Very basic - Factual, brief and concise	- More details and precision - Indicate time, direction, and speed appropriately - Road sign labelling

REFERENCES

[1] M. Gadd, D. De Martini, L. Marchegiani, P. Newman, and L. Kunze, "Sense–assess–explain (sax): Building trust in autonomous vehicles in challenging real-world driving scenarios," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 150–155, IEEE.

[2] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[3] 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.

[4] N. T. S. Board, "Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator mountain view, california," "(accessed October 30, 2020)".

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, IEEE, 2018.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[7] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference," *arXiv preprint arXiv:2003.02425*, 2020.

[8] L. Kunze, T. Bruls, T. Suleymanov, and P. Newman, "Reading between the lanes: Road layout reconstruction from partially segmented scenes," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 401–408, IEEE, 2018.

[9] H. H. Kelley, "The processes of causal attribution.," *American psychologist*, vol. 28, no. 2, p. 107, 1973.

[10] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *Proceedings of the 2018 Chi conference on human factors in computing systems*, pp. 1–14, 2018.

[11] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2119–2128, 2009.

[12] 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.

[13] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.

[14] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9523–9532, 2020.

[15] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *arXiv preprint arXiv:2103.05154*, 2021.

[16] 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.

[17] 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 (IJIDeM)*, vol. 9, no. 4, pp. 269–275, 2015.

[18] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[19] S. Sreedharan, T. Chakraborti, and S. Kambhampati, "Balancing explicability and explanation in human-aware planning," in *2017 AAAI Fall Symposium*, pp. 61–68, AI Access Foundation, 2017.

[20] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, "Ai challenges in human-robot cognitive teaming," *arXiv preprint arXiv:1707.04775*, 2017.

[21] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, pp. 9505–9515, 2018.

[22] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.

[23] A. F. Winfield, K. Winkle, H. Webb, U. Lyngs, M. Jirotka, and C. Macrae, "Robot accident investigation: a case study in responsible robotics," *arXiv preprint arXiv:2005.07474*, 2020.

[24] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

[25] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Visual explanation by attention branch network for end-to-end learning-based self-driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1577–1582, IEEE, 2019.

[26] A. F. Winfield and M. Jirotka, "The case for an ethical black box," in *Annual Conference Towards Autonomous Robotic Systems*, pp. 262–273, Springer, 2017.

[27] 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.

[28] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[29] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.