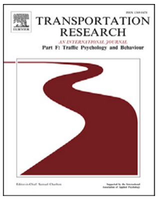
Transportation Research Part F: Psychology and Behaviour 109 (2025) 1275–1292

Contents lists available at ScienceDirect

Transportation Research Part F: Psychology and Behaviour

journal homepage: www.elsevier.com/locate/trf



A transparency paradox? Investigating the impact of explanation specificity and autonomous vehicle imperfect detection capabilities on passengers

Daniel Omeiza [a,*], Raunak Bhattacharyya [a,c,1], Marina Jirotka [a], Nick Hawes [a], Lars Kunze [a,b]

[a] *University of Oxford, Oxford, United Kingdom*
[b] *University of the West of England, Bristol, United Kingdom*
[c] *Indian Institute of Technology, Delhi, India*

ARTICLE INFO

Keywords:
Explanations
Transparency
Autonomous driving
Perceived safety
Visual attention

ABSTRACT

Transparency in automated systems could be afforded through the provision of intelligible explanations. While transparency is desirable, might it lead to catastrophic outcomes (such as anxiety) that could outweigh its benefits? It's quite unclear how the specificity of explanations (level of transparency) influences recipients, especially in autonomous driving (AD). In this work, we examined the effects of transparency mediated through varying levels of explanation specificity in AD. We first extended a data-driven explainer model by adding a rule-based option for explanation generation in AD and then conducted a within-subject lab study with 39 participants in an immersive driving simulator to study the effect of the resulting explanations. Specifically, our investigation focused on: (1) how different types of explanations (specific vs. abstract) affect passengers' perceived safety, anxiety, and willingness to take control of the vehicle when the vehicle perception system makes erroneous predictions; and (2) the relationship between passengers' behavioural cues and their feelings during the autonomous drives. Our findings showed that abstract explanations did not make passengers safer despite being vague enough to conceal all perception system detection errors compared to specific explanations having a minimal amount of exposed perception system detection errors. Anxiety levels increased when specific explanations revealed perception system detection errors (high transparency). We found no significant link between passengers' visual patterns and their anxiety levels. We advocate for explanation systems in autonomous vehicles (AV) that can adapt to different stakeholders' transparency needs.

1. Introduction

The automotive industry has witnessed an increasing level of development in the past decades, from manufacturing manually operated vehicles to manufacturing vehicles with a high level of automation. Despite these technological strides, accidents involving

---

* Corresponding author.

E-mail addresses: danielomeiza@robots.ox.ac.uk (D. Omeiza), raunak@robots.ox.ac.uk (R. Bhattacharyya), marina.jirotka@ox.ac.uk (M. Jirotka), nickh@robots.ox.ac.uk (M. Hawes).

[1] This author was with the University of Oxford at the time this work was done. This author's affiliation is now the Indian Institute of Technology, Delhi.

https://doi.org/10.1016/j.trf.2025.01.015
Received 17 July 2024; Received in revised form 10 January 2025; Accepted 14 January 2025

vehicles operating in autonomous mode continue to undermine public trust (Stanton et al., 2019; Board, 2018; Lavrinc, 2016; Tilley, 2016; McFarland, 2016), irrespective of who is at fault. As highly automated vehicles make high-stake decisions that can significantly affect passengers, the vehicles should explain or justify their decisions to meet set transparency guidelines or regulations, e.g., GDPR Article 12 (Voigt & dem Bussche A, 2017) and the (Information Commissioner's Office, 2021).

In this paper, autonomous vehicle (AV) refers to SAE Level 4 vehicles (SAE, 2021) that operate without human inputs in geofenced environments. Accompanying AV driving decisions with natural language explanations is one promising approach for better vehicle transparency (Omeiza et al., 2022b; Ha et al., 2020; Koo et al., 2015). This transparency, obtained through intelligible explanations, can help to reassure passengers of safety and also assist them in effectively calibrating their trust in an AV (Khastgir et al., 2018). The specificity level of explanations is, however, an important factor in achieving the aforementioned benefits. In real-world deployments, AVs may not always achieve perfect scene understanding due to the limitations of their perception systems. Depending on the specificity level, explanations might reflect these flaws even when they are inconsequential. Informing operators about these inherent imperfections could enhance safety by helping them recognise when to remotely take control of the vehicle (Kunze et al., 2019a). Explanation logs (aligned with driving scenes) could also be helpful to AV developers in improving the perception system performance. While this transparency is helpful for these groups, it remains uncertain whether in-vehicle passengers would prefer high-level transparency (if at all useful) that reveals such minor errors (such as mistaking a van for a bus). This is what we refer to as the transparency paradox in this paper. It is, therefore, important to determine the appropriate level of transparency for in-vehicle passengers, mediated through the specificity of the explanations.

Furthermore, as passengers are likely to engage in other activities during their ride, relying solely on visual cues to communicate awareness may be ineffective when passengers' attention is desired. Hence, auditory feedback (Kunze et al., 2019b) is also necessary to ensure passengers are adequately informed.

In this study, we investigate the effects of explanation specificity on AV passengers' perceived safety, feeling of anxiety, and desire to takeover control. We test two levels of explanation specificity: abstract and specific. We use the term *abstract* to describe the provision of vague auditory explanations that conceal details about a driving situation (including perception system detection errors). In contrast, *specific* refers to the provision of very detailed and fine-grained explanations about a situation.

1.1. Contribution statement

Overall, this research makes three contributions to the fields of explainable autonomous driving and human-machine interaction:

1. It sets out a new case study of explanation specificity in the presence of perception systems errors in the autonomous driving context;
2. It provides an enhanced interpretable technique for generating textual and auditory natural language explanations for AV navigation actions;
3. It reveals experimental findings on whether high AV transparency, though critical to other stakeholders, is helpful to AV passengers.

2. Background

This section provides a background on explanations for automated decisions and in autonomous driving from the literature.

2.1. Explanations for automated decisions

The concept of explanations has been studied extensively by a multidisciplinary group of scholars, ranging from Philosophy to Psychology. Each adopts an idiosyncratic lens, characterising explanations in terms relevant to the goals of their respective disciplines. Therefore, to guide our study, we align with a definition proposed in an extensive survey of explanations in autonomous driving (Omeiza et al., 2022b): explanations are a piece of information presented in an intelligible way as a reason or part of a reason for an outcome, event or an effect in text, speech, or visual forms.

Recent efforts around explanations have been mainly channelled towards complex AI systems (explainable AI (Gunning, 2017; Adadi and Berrada, 2018; Došilović et al., 2018)) to understand and communicate the reasons for the systems' decisions. Techniques developed for this purpose fall under different categories based on their mode of operation. Some are model-specific (Chakraborti et al., 2020) in that they investigate the underlying AI algorithm in detail to support debugging tasks (Magnaguagno et al., 2017). These types of explainers are intrinsic and model specific meaning that they are inherently coupled with the underlying algorithm. Other explanation approaches are classified as model-agnostic as they can assess the properties of an output independent of the algorithms used to realise the output (Ribeiro et al., 2016; Lundberg & Lee, 2017). These explainers aim to help enhance user's knowledge of a system and potentially foster trust.

Explanations must possess certain properties to be effective to their intended recipients. Mittelstadt et al. (2019) 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 social, selective, and contrastive explanations. Social in the sense that the explanation process involves different parties and the explainer is able to model the expectations of the explainee. The explanation is selective if it can select explanations from several competing hypotheses. It is considered contrastive if it can differentiate the properties of two competing hypotheses. Kment (2006) further emphasised the value of counterfactual explanations in enhancing understanding.

These explanations describe how changes in input can lead to a shift from one fact to a competing hypothesis (foil) (Miller, 2019). Explanations should be intelligible (Omeiza et al., 2021b) and strike a delicate balance between providing sufficient detail and respecting the cognitive capacity of the explainee. This balance presents a significant challenge, as accurately gauging an individual's real-time cognitive capacity remains difficult.

Different methodologies have been adopted in the XAI literature. Wang et al. (2019) categorised these research methodologies into three groups: First, the existence of *unvalidated guidelines* for the design and evaluations of explanations was highlighted. The authors claimed that these kinds of guidelines are based on authors' experiences with no further substantial justification. Second, researchers suggested (in Zhu et al. (2018)) that understanding users' requirements is helpful in XAI research. It is on this premise that previous research on explanation design has been thought to be *empirically derived*. This type of XAI research elicits explanation requirements from user surveys to determine the right explanation for a use case with explanation interfaces. Third, some explanation design methods are derived from *psychological constructs from formal theories* in the academic literature. Some of these methods (e.g., in Hoffman and Klein (2017)) draw on theories from cognitive psychology to inform explanation design for explanation frameworks. Our work is heavily grounded on empirical studies.

2.2. Explanations in autonomous driving

Explanations have been found useful in enhancing user experience (Schneider et al., 2021), trust (Koo et al., 2015; Ha et al., 2020), and improved situational awareness (Omeiza et al., 2021a; Liu et al., 2021) in automated driving. Recent works have explored human factors in the application of explainable AI in AD. For instance, in Omeiza et al. (2021b, 2022a), a socio-technical approach to explainability was proposed. An interpretable representation and algorithms for explanations based on a combination of actions, observations, and road rules were designed. Regarding explanations depths, the notion that explanations with higher levels of abstraction and correctness are superior has been argued in the literature (Buijsman, 2022; Guidotti et al., 2018). Additionally, Ramon et al. (2021) argued that the specificity of explanations should be tailored to the application context, noting that low-level specificity is often preferred by individuals with a more deliberative cognitive style.

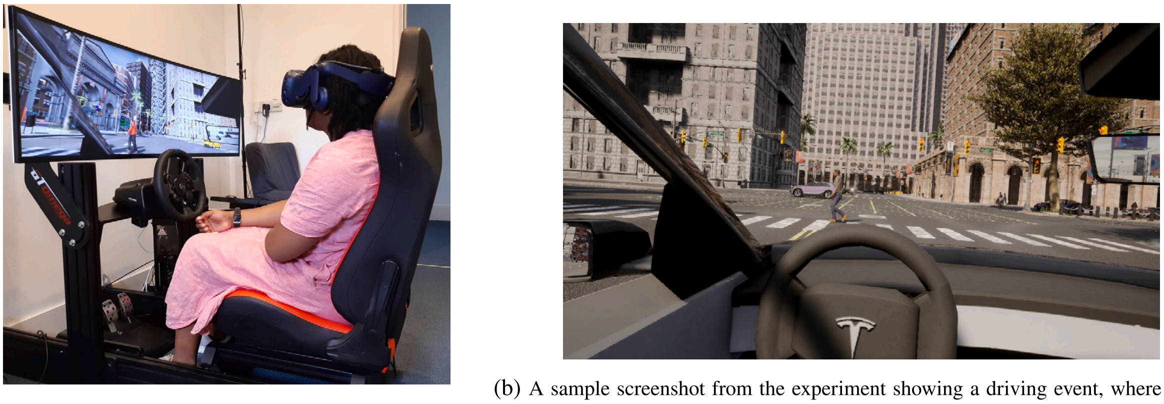
In this paper, the term *explanation specificity* is used to refer to two specificity levels of explanations, *abstract* (low transparency) and *specific* (high transparency). Explanations can be used to convey different information in AD, e.g., vehicle uncertainties and intentions, and communicated through different modalities. For example, Kunze et al. (2019a) conveyed visual uncertainties with multiple levels to operators using heartbeat animation. This information helped operators calibrate their trust in automation and increased their situation awareness. Similarly, Kunze et al. (2019b) used peripheral awareness display to communicate uncertainties to alleviate the workload on operators simultaneously observing the instrument cluster and focusing on the road. This uncertainty communication style decreased workload and improved takeover performance. In addition, the effects of augmented reality visualisation methods on trust, situation awareness, and cognitive load have been investigated in previous studies using semantic segmentation (Colley et al., 2021), scene detection and prediction (Colley et al., 2022), pedestrian detection and prediction (Colley et al., 2020), situation detection and prediction for crowdsourced videos of challenging driving scenes (Jansen et al., 2024). These deep vision-based techniques applied to automated driving videos and rendered in augmented reality mode were a way of calling the attention of operators to risky traffic agents in order to enhance safety. Our work builds on the existing works by exploring a new direction with respect to natural language explanation specificity rendered in auditory form. The existing works have mainly focused on visual signals, leaving auditory natural language explanations under-explored, especially for high automation levels (e.g., Level 4 and 5) where human inputs are not required. Auditory means of communicating explanations (Kaufman et al., 2024; Nees et al., 2016) are important for situational awareness, especially at the early deployment stages of AV in society, where scepticism is high. We thus used an auditory communication style in this study to convey natural language explanations to passengers. Natural language explanations were constructed following the explanation template used in Omeiza et al. (2022a) where actor types, location, actions, and reasons for actions are factored in.

Some existing works around human-machine interaction (Liu et al., 2021) have leveraged theoretical models (e.g., mental and situational models (Endsley, 2000)) to study explanations. We based our work on behavioural cues and subjective feedback from participants.

2.3. Research questions

From the preceding literature review, we find the need to gain a deeper understanding of the effects of transparency, brought about by natural language explanations of varying specificity, especially under imperfect AV perception systems.

1. Given varying levels of perception system detection errors, how do natural language explanations influence passengers' perceived safety? The perception system uses a variety of sensors to sense and understand its environment. Perception system error occurs when the system mis-interprets its environment. In this paper, an error refers to the misclassification of actors in a driving scene.
    - **H1.1 - Perceived Safety.** *Low transparency yields a higher perception of safety in an AV with perception system detection errors.* We hypothesise that passengers feel safer in a low transparency AV, despite receiving abstract explanations. While individuals often seek the truth, many prefer information that aligns with their expectations (Hart et al., 2009). Consequently, specific explanations may reveal perception system detection errors that contradict passenger expectations. Additionally, research has shown that placebo explanations can have similar positive effects on people as real explanations (Eiband et al., 2019a).
    - **H1.2 - Feeling of Anxiety.** *Passengers' feeling of anxiety increases with increasing perception system detection errors in a highly transparent AV.* We posit that there is a connection between perceived safety and the feeling of anxiety (Davidson et al., 2016;



(a) Driving simulation setup for the study.

(b) A sample screenshot from the experiment showing a driving event, where the AV had to stop at red light, and a pedestrian crossing.

Fig. 1. Driving simulation setup for the study. The setup included a VR headset, steering wheel, brake and acceleration pedals, screen, and arcade seat. The participants experienced the scenarios in a VR environment. The screen helped the researcher observe the experiment. In the captured scene in Fig. 1b, a pedestrian crosses at the crosswalk.

Quansah et al., 2022). Therefore, explanations that frequently reference misclassified actors are likely to create a sense of insecurity, leading to increased anxiety.

- **H1.3 - Takeover Feeling.** *In highly transparent AVs, passengers are more likely to develop the feeling to take over navigation control from the AV with higher perception system detection errors.* Although passengers are not able to take control in this study, we anticipated that they might nurse the thought to do so if they repeatedly received illogical explanations from the AV.

2. Do passengers' behavioural cues correlate with their feelings?
   - **H2.1 - Visual Feedback** *Visual feedback from participants correlates with their feeling of anxiety.* Individuals with the feeling of anxiety might be usually hyper-aroused and sensitive to environmental stimuli. They may have difficulties concentrating, performing tasks efficiently, and inhibiting unwanted thoughts and distractions (Hepsomali et al., 2017; Chen et al., 2014). Participants' fixation points and saccades should correlate with anxiety.

3. Passenger study

In this section, we describe the participants' demographic, experiment apparatus setup, experiment design, and the procedure of the experiment. The necessary approval to conduct the study was obtained from our University's Research Ethics Committee.

3.1. Participants

We conducted a power analysis (with a medium effect size) to inform our decision regarding the number of subjects. Afterwards, calls for participants were placed on various online platforms, such as the call for participants platform (Call For Participants, 2022), university mailing groups, university Slack channels, the research group website, and social media to recruit subjects. Upon screening, the final sample consisted of $N = 39$ participants (28 male, 11 female) ranging in age from 18 to 59 years. The participants comprised students, university employees, and members of the callforparticipants platform. Although prior driving experiences were not required, 28 (71.79%) of the participants were licensed drivers. Only 2 of the 39 participants (5.13%) had experience with autonomous drives, however, in a research context. 6 (15.38%) of the participants had used a virtual reality headset for a driving game or driving experiment in the past.

3.2. Apparatus

3.2.1. Hardware

The hardware setup is shown in Fig. 1. We conducted the experiment in a driving simulator that comprised a GTR arcade seat, Logitech G29 steering wheel with force feedback, turn signal paddles, brake and accelerator pedals, and an ultra-wide LG curved screen to display the experiment. A state-of-the-art virtual reality (VR) headset (with an immersive 360° FoV and an eye tracker) was also used to provide an immersive experience and high visual fidelity.

3.2.2. Driving software

Software architecture is illustrated in Fig. 2. We adapted the DReyeVR (Silvera et al., 2022), an open-source VR-based driving simulation platform for behavioural and interaction research involving human drivers. DReyeVR was built atop CARLA (Dosovitskiy et al., 2017), an open-source driving simulator for AD and Unreal Engine 4. DReyeVR provides a very realistic experience with naturalistic visuals (e.g., in-vehicle mirrors) and auditory (e.g. vehicular and ambient sounds) interfaces allowing for an ecologically valid setup. It also provides an experimental monitoring and logging system to record and replay scenarios, as well as a sign-based navigation system.

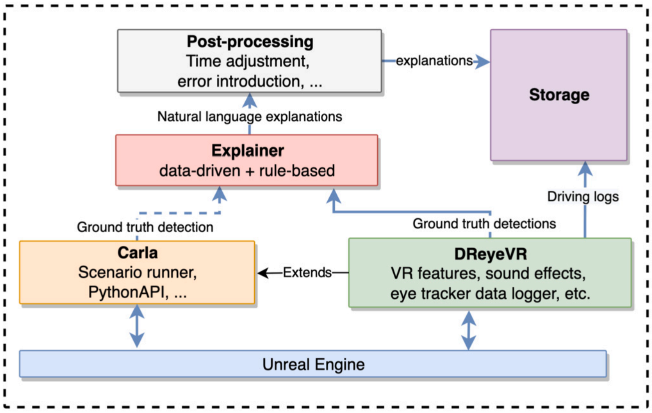


Fig. 2. High-level architecture of our simulation software. DReyeVR uses Unreal engine and extends CARLA simulator, which also builds on Unreal engine. DReyeVR extends CARLA by adding VR functionalities, vehicular and ambience sounds, eye tracker data logging, and additional sensors, among others. Our explainer model, which is both rule-based and data-driven, receives ground truth data from CARLA or DReyeVR and generates explanations for predicted actions. The post-processing script allowed us to modify the generated explanations as we desire.

3.2.3. *Explainer software*

As shown in Fig. 2, we developed an explainer system based on previous work (Omeiza et al., 2022a). This system utilises a tree-based model trained on an annotated AV driving dataset that we collected in a prior project. While the original algorithm in Omeiza et al. (2022a) is primarily data-driven, we incorporated a rule-based technique to serve as a fallback when the data-driven method fails or makes an incorrect ego action prediction. The data-driven method employs a trained tree-based model to predict and generate explanations from detections obtained from CARLA, a driving simulator. In contrast, the rule-based approach relies on CARLA's ground truth data and follows predefined rules to determine which agents to reference in the explanations. By comparing predictions from the data-driven method with ground truth observations from CARLA, we can identify incorrect predictions. This enhanced explainer system, combining both data-driven and rule-based approaches, was used to generate preliminary explanations for our created scenarios. Wintersberger et al. (2020) suggested types of traffic elements to be included in visual explanations based on user preferences. Our proposed explainer, however, selects traffic elements deemed important (feature importance Anjomshoae et al., 2021) by the driving model for its decisions (see Algorithm 1).

We performed post-processing operations on the generated explanations, including fine-tuning some of the content and adjusting timestamps to ensure the explanations were delivered at the appropriate moments.

Algorithm 1: Intelligible Explanation Generation.

```
    Input: tree model $\mathcal{M}$ for ego's action prediction, input vector $\mathbf{X}$ describing ego's environment
    Output: intelligible auditory explanation
1   Select a representative tree $m \in \mathcal{M}$ from tree model $\mathcal{M}$.
2   Predict action $y \in \mathcal{Y}$ given $\mathbf{X}$.
3   Compare prediction $y$ with CARLA ground truth $y_{GT}$.
4   if $y = y_{GT}$ then
5       Trace the decision path $\mathcal{P}_y$ for the prediction $y$ in tree $m$.
6       Compute the importance score $I(\mathbf{X}_i)$ of the attributes $\mathbf{X}_i$ in each node $i$ along the decision path $\mathcal{P}_y$.
7       Select attributes $\mathbf{X}_i$ with importance scores $I(\mathbf{X}_i) \geq k$, where $k$ is a predefined threshold.
8       Merge the conditions/inequalities in the selected attributes $\mathbf{X}_i$.
9       Translate merged attributes $\mathbf{X}_i$ to natural language following the template in Table 2.
10  else
11      Use CARLA ground truth information $y_{GT}$ and predefined rules to generate explanation following the template in Table 2.
```

To estimate feature importance scores (line 6), we evaluated tree-specific methods, including local increments (Palczewska et al., 2013) and Tree SHAP (Lundberg & Lee, 2017), to identify a suitable approach. We selected the Tree SHAP algorithm for its superior accuracy. Tree SHAP calculates SHAP values for each feature. The core procedure is as follows:

- generate all subsets S of the set $F = \{1, ..., N \setminus \{i\}\}$
- for each $S \subseteq F \setminus \{i\}$ estimate the contribution of feature i as $CT\{i|S\} = \mathcal{M}(S \cup \{i\}) - \mathcal{M}(S)$
- compute the SHAP value according to:

$$\phi_i := \frac{1}{N} \sum_{S \subseteq F \setminus \{i\}} \binom{N-1}{|S|}^{-1} CT(i|S) \tag{1}$$

Table 1
Description of Scenarios: The Independent Variables. Note that 'Perception Errors' refers to the percentage of explanations that contain errors for the particular drive. Perception errors are not obvious (0%) when the explanations are vague/abstract.

Scenario	Specificity/Transparency	Perception Errors (%)	Scenario Length	Examples of Events	Environment
Abstract	Low	0	~ 4 minutes	All events from Table 1	CARLA Town10HD
Specific (5)	High	5	~ 4 minutes	All events from Table 1	CARLA Town10HD
Specific (50)	High	50	~ 4 minutes	All events from Table 1	CARLA Town10HD

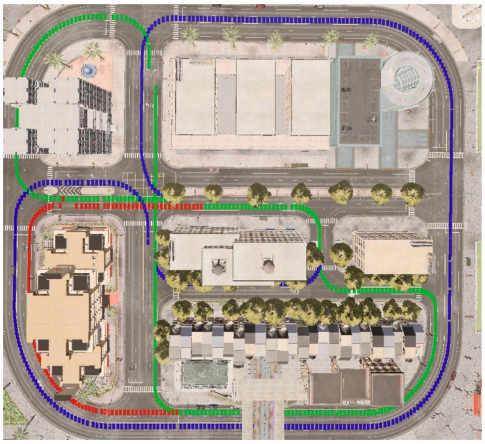


Fig. 3. Scenario routes. **Red**: Abstract, **Green**: Specific(5), **Blue**: Specific(50). Each route is a loop and overlaps with others at some points. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

High positive SHAP values signify high importance and utility, while very low negative SHAP values indicate high importance but low utility.

4. Experiment design

Before the start of the trials, participants were asked to manually drive a vehicle for about two minutes in CARLA Town03—a complex town, with a 5-lane junction, a roundabout, unevenness, and a tunnel. 30 vehicles and 10 pedestrians were spawned in this town. This preliminary drive aimed to familiarise participants with the driving simulation environment and to allow them to experience manual driving within the simulation.

We employed a within-subject design due to the limited sample size, which was insufficient for a between-subject study. Additionally, this design helped mitigate the potential co-founding factor of between-individual differences in a between-subject design.

4.1. Independent variable

Combinations of transparency level (Low and High) and AV perception system detection errors (low and high) were done to obtain the independent variable *Scenarios*. The first scenario (*Abstract* scenario) comprises abstract explanations indicating low transparency, and although with an imperfect perception system, perception system detection errors are not obvious (0%) due to low transparency. The second scenario (*Specific(5)* scenario) comprises specific explanations indicating high transparency and 5% amount of perception system detection errors indicating low error degree. The third scenario (*Specific(50)* scenario) comprises specific explanations indicating high transparency and 50% amount of perception system detection errors indicating high error degree. The driving events that made up the different scenarios were carefully designed to include different driving conditions that are obtainable in the real world (See Table 2).

The scenario routes are shown in Fig. 3.

i. Abstract A scenario in CARLA Town10HD, which is about 4 minutes long (330 secs). Town10HD is an urban city environment with different infrastructures, such as an avenue or promenade, and realistic textures. Driving conditions are a combination of the events in Table 2. The perception system in this scenario might contain some errors, but the explanations provided in this scenario were post-processed to always provide surface information which is vague enough to conceal perception system detection errors. The rules governing explanations for this scenario were:

- all traffic lights are referred to as 'traffic sign' without specifying the state (e.g., red, green, amber, off) of the traffic light;
- pedestrians are referred to as 'road users';
- all non-human moving actors are referred to as 'vehicle'. This includes cycles, motorbikes, cars, etc.

Table 2
Description of a subset of the events (5 out of 9) and corresponding explanations provided during the study. Observations and causal explanations are announced to passengers. AV's action (text in red), other agent's class & action (text in blue), and the agent's location (text in green) are determined by the explainer algorithm described in Algorithm 1. Note that the specificity of explanations and observations is set by modifying the class of the external actor/agent, e.g., 'motorbike' is specific, while 'vehicle' is abstract as it doesn't say the vehicle type. We have classified observations as explanations in this paper. The only difference between causal explanations and observations is that the earlier provides a cause for a driving action while the latter provides information about observed actors while the AV is not executing a new action.

Event	Description	Observation	Causal Explanation
FollowLeadingVehicle	AV follows a leading actor. At some point, the leading actor slows down and finally stops. The AV has to react accordingly to avoid a collision.	vehicle ahead on my lane.	Stopping because cyclist stopped on my lane.
VehicleTurning	AV takes a right or a left turn from an intersection where an actor suddenly drives into the way of the AV, AV stops accordingly. After some time, the actor clears the road, AV continues driving.	motorbike crossing my lane.	Stopping because motorbike is crossing my lane.
LaneChangeObstacle	AV follows a leading actor, and at some point, the leading actor decelerates. The AV reacts accordingly by indicating and then changing lanes.	vehicle ahead on my lane.	Changing lane to the [right/left] because vehicle stopped on my lane.
StopSignalNoActor	No actor ahead of the AV at a signalised intersection with a red traffic signal. AV decelerates and stops.	red traffic light ahead on my lane.	Stopping because traffic light is red on my lane.
MovSignalNoActor	No actor ahead of the AV. AV starts moving from a stop state at a signalised junction or intersection.	None	Moving because traffic light is green on my lane.

An example explanation is 'stopping because of the traffic sign on my lane'. This obfuscates the type and colour of the traffic sign.

ii. Specific(5) A scenario in CARLA Town10HD, which was about 4 minutes in length (256 seconds). Driving conditions in this scenario were a combination of the events in Table 2. The explanations generated in this scenario were specific and detailed, exposing all errors. The perception system of the AV in this scenario was about 5% inaccurate. This error value was estimated following the dynamic traffic agent classification model and confusion matrix provided by Bin Issa et al. (2021) and the traffic light classification model and confusion matrix by Michael & Schlipsing (2015). We were only interested in the confusion matrices (and not the models). The confusion matrices helped us to systematically introduce the 5% perception system detection errors during the post-processing stage of the explanations. In this scenario, the 5% error resulted in one explanation (1 out of the 22) being erroneous as the explanation exposed the misclassification errors from the perception system. An example of an erroneous explanation is: 'van ahead on my lane'. Here, a car was misclassified as a van.

iii. Specific(50) A scenario in CARLA Town10HD, which was 4 minutes in length (274 seconds). Driving conditions were a combination of the events in Table 2. The explanations generated in this scenario were as fine-grained/specific and detailed as those in the *Specific(5)* scenario. The perception system error of the AV in scenario *Specific(5)* was significantly noised to reach a reduced accuracy of 50%. We assumed that this reduction in accuracy might be sufficient to influence peoples' behaviour. Therefore, half of the explanations in this scenario (12 out of 24) reflected misclassification of actors or actor states. An example of an erroneous explanation is 'moving because traffic light is switched off on my lane'. In this case, the perception system failed to identify a green light accurately.

We chose only two scenarios for specific explanations because we only wanted to simulate a realistic system (Specific(5)) and an unrealistic system, one with high detection error (Specific(50)). We established that these two conditions were sufficient to investigate the impact of an erroneous perception system propagated through explanations. Also, note that all three scenarios were designed so that the AV perception system detection errors were insignificant to the AV's navigation actions. Hence, the AV respected all road rules and avoided collisions. This was important as the state-of-the-art AVs would likely not make obvious navigation errors. Moreover, we were interested in the effects of the awareness of inconsequential detection errors in AVs. Hence, it was necessary to introduce artificial detection errors of varying degrees (low and high). Perception errors, in this case, that define Specific(5) and Specific(50) refer to the percentage of explanations that contain obvious errors in a particular drive. Perception errors are not obvious (0%) when the explanations are abstract/vague (a way to describe lack of transparency/low transparency). The non-influence of AV perception system detection errors on navigation control also helped to avoid the confounding factors of route navigation problems. Further, we counterbalanced the routes across scenarios. That is, the AV's route was different in each scenario. This design decision was made to reduce carry-over effects on the participants. With this setup, the scenarios were still comparable as they were all within the same town, and the routes shared similar features. Each scenario also had a balanced combination of the events listed in Table 2. In all the scenarios, the AV maintained a speed below $30mph$, the recommended speed limit in urban areas in the UK. See Fig. 4 for sample scenes from each scenario and their corresponding explanations.

4.2. Dependent variables

There were six dependent variables: *Perceived Safety, Feeling of Anxiety, Takeover Feeling, Fixation Divergence, Saccade Difference,* and *Button Presses*. These variables were categorised into two (psychological factors and behavioural cues) for easy analysis and reporting.

(a) Abstract - *Observation:* **'vehicle is crossing on my lane'**

(b) Specific(5) - *Observation:* **'motor bike ahead on my lane'**

(c) Specific(50) - *Causal Explanation:* **'stopping because cyclist is crossing my lane.'**

Fig. 4. Sample screenshots and the generated explanations (including observations announcement and causal explanations) from the three driving scenarios. Heatmaps of gaze points from all the participants are plotted over the images, indicating areas of interest. In the *Abstract* scenario (Fig. 4a), all movable/dynamic non-human actors are referred to as 'Vehicle'. Thus, a cyclist was referred to as a vehicle. Fig. 4b depicts a scene from the *Specific(5)* scenario in which the AV's perception system accurately identified and classified a motorbike and provided a fine-grained explanation for this. In the *Specific(50)* scene (Fig. 4c), the AV's perception system misclassified a pedestrian as a cyclist. The fine-grained/specific explanation provided exposed this error.

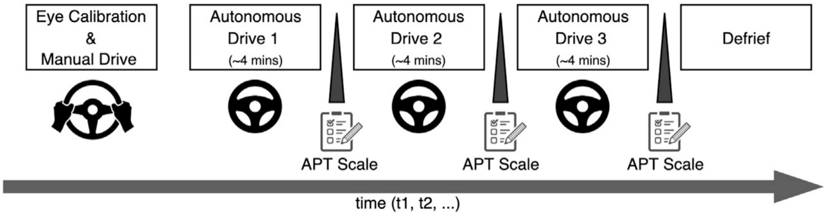


Fig. 5. Study procedure. Eye calibration was done with the VR headset; participants drove for two minutes, participants experienced each of the 4 mins scenarios in counterbalanced order and completed the Feeling of Anxiety, Perceived Safety, and Takeover Feeling Questionnaire (APT Scale) in between each scenario. Participants were debriefed.

Psychological factors These factors include *Perceived Safety, Feeling of Anxiety*, and *Takeover Feeling*. They were mainly measured using items from the Autonomous Vehicle Acceptance Model Questionnaire (AVAM) (Hewitt et al., 2019). AVAM is a user acceptance model for autonomous vehicles, adapted from existing user acceptance models for generic technologies. It comprises a 26-item questionnaire on a 7-point Likert scale, developed after a survey conducted to evaluate six different autonomy scenarios.

Items 24—26 were used to assess the *Perceived Safety* factor, while items 19—21 were used to assess the *Feeling of Anxiety* factor. Similar to Schneider et al. (2021), we introduced a new item to assess participants' feelings to takeover navigation control from the AV during the ride (*Takeover Feeling*). Specifically, participants were asked to rate the statement 'During the ride, I had the feeling to take over control from the vehicle' on a 7-point Likert scale. Actual navigation takeover by participants was not permitted because we wanted to be able to control the entire experiment and have all participants experience the same scenarios. Moreover, we were dealing with L4 automation. Though participants were not expected to drive or take over control, they might have nursed the thought to do so. This is what the *Takeover Feeling* variable measures.

We added a free-response question related to explanations with the aim of obtaining qualitative data for triangulating quantitative results. Participants were asked the following question: 'What is your thought on the explanations provided by the vehicle, e.g., made you less/more anxious, safe, feeling to take over control?'. We refer to the resulting questionnaire as the APT Questionnaire (i.e., A-Anxiety, P-Perceived Safety, T-Takeover Feeling).

Behavioural cues We also used *Button Presses, Fixation Divergence*, and *Saccade Difference* as additional metrics. *Button Presses* on the steering wheel were measured when participants felt unsafe, anxious, or confused feelings.

Fixation Divergence is the Euclidean distance between mean participants' fixation points and reference fixation points. This provides information to draw inferences about participants' distractions.

For *Saccade Difference*, we estimated participants' saccade velocity over time following the method in Gibaldi and Sabatini (2021) and found the difference from a reference saccade velocities. Saccade is the rapid movement of the eye between fixation points. Saccade velocity is the speed of such movements. The fixation and saccade reference points (or ground truths) were the fixation and saccade records obtained from the researcher, who also participated in the study.

4.3. Procedure

The procedure of the experiment is illustrated in Fig. 5. After all preliminary form completions and briefings, we introduced the physical driving simulator and explained the subsequent steps, which involved a pre-experiment manual driving session in VR mode lasting for 2 minutes. Participants were informed that the purpose of this pre-experiment exercise was to familiarise them with the simulation environment and to identify individuals prone to motion sickness, who would then be excluded from the main experiment.

Table 3
Descriptive statistics from APT questionnaire analysis.

	Perceived Safety Cronbach α : 0.87, $H(2) = 8.17, \boldsymbol{p = .017}$			Feeling of Anxiety Cronbach α : 0.86 $H(2) = 13.32, \boldsymbol{p = .001}$			Takeover Feeling $H(2) = 6.27, \boldsymbol{p = .044}$		
	Mean	SD	Mean Rank	Mean	SD	Mean Rank	Mean	SD	Mean Rank
Abstract	4.89	1.35	2.15	2.81	1.34	1.72	2.79	1.91	1.68
Specific(5)	**4.93**	1.13	**2.22**	2.79	1.2	1.81	3.31	1.79	2.10
Specific(50)	3.86	1.58	1.63	**3.93**	1.68	**2.47**	**3.87**	1.94	**2.22**

Upon completion of the manual driving exercise, the researcher removed the VR headset from the participant and explained the aim and procedure of the main experiment. The instructions included the following statements: 'you would experience 3 autonomous rides by different vehicles, [...] and after each ride, you would complete a short survey. The vehicle drives along a predefined path for about 4 minutes and provides explanations for its planned driving decisions and announces relevant objects in its environment. [...]. The vehicle tells you its next direction at a junction or an intersection using its right or left red light indicators on its dashboard accordingly.[...] Simply click any of these buttons if the decision or the explanation of the vehicle makes you feel confused, anxious or unsafe [...]'. The researcher then placed the VR headset back on the participant and initiated the scenarios. Complete counterbalancing was applied to the scenario treatments, resulting in six different orders of scenarios. Each participant experienced the scenarios in one of these six orders, with approximately six participants per order.

Participants were encouraged to rest briefly after each driving experience, with the VR headset removed. A short debriefing session followed the study, and participants were given a £10 Amazon gift card. The entire experiment lasted approximately 50 minutes.

The researcher also participated in the experiment, experiencing all three scenarios. Throughout, the researcher focused on the lane ahead and the actors referenced by the explanations. Neither erroneous nor abstract explanations influenced the researcher's focus, as the researcher remained attentive to the lane and the actors or obstacles impacting the AV's actions, regardless of the explanations. This consistency was due to the researcher's familiarity with all scenarios. The data from the researcher served as a reference or ground truth. Notably, the researcher's eye movements matched normal human saccadic velocity, which reaches 300—400°/seconds (Raab, 1985; Wilson et al., 1993).

5. Quantitative results

5.1. Psychological factors analysis

To test our hypotheses listed in Section 2.3, we analyzed data from the three APT questionnaires. We created a latent variable (*Feeling of Anxiety*) by averaging responses to AVAM Items 19—21, and another latent variable (*Perceived Safety*) by averaging responses to AVAM Items 24—26. We calculated Cronbach's Alpha (α) for the independent variables that formed the latent dependent variables to ensure they had adequate internal consistency. Results with an adjusted p-value less than 0.05 ($p < .05$) were considered significant. P-values were adjusted using Bonferroni corrections, where the calculated p-values were multiplied by the number of scenarios, to reduce the likelihood of Type I errors (false positives). Normality tests, including the Kolmogorov-Smirnov, Shapiro-Wilk, and Anderson-Darling tests, indicated a violation of normality in the *Feeling of Anxiety*, *Perceived Safety*, and *Takeover Feeling* factors. Hence, we performed a Friedman test for these dependent variables, see Table 3 and Fig. 6.

H1.1 - perceived safety *Low transparency yields a higher perception of safety in an AV with perception system detection errors.*

A Friedman test was conducted. No significant difference was found in the scenario pair: *Abstract — Specific(5)*, and the pair: *Abstract — Specific(50)*. In fact, the *perceived safety* mean rank in the *Specific(5)* scenario (2.22) was higher than that in the *Abstract* scenario (2.15), see Table 3. Therefore, there was no sufficient evidence in support of hypothesis H1.1.

However, we observed a statistically significant difference in perceived safety between *Specific(5)* and *Specific(50)* scenarios $p = .017$.

H1.2 - feeling of anxiety *Passengers' feeling of anxiety increases with increasing perception system detection errors in a highly transparent AV.* A Friedman test indicated a significant difference in the *Feeling of Anxiety* across scenarios, $H(2) = 13.32, p = .001$. The pairwise scenario comparisons of *Abstract - Specific(50)* and *Specific(5) - Specific(50)* resulted in an adjusted p-value of .003 and .01 respectively (see Table 3). Hence, there is strong evidence in support of hypothesis H1.2.

H1.3 - takeover feeling *In highly transparent AVs, passengers are more likely to develop the feeling to take over navigation control from the AV with higher perception system detection errors.* A Friedman test showed a significant difference in *Takeover Feeling* across scenarios, $H(2) = 6.27, p = .044$. While the pairwise scenario comparison of *Abstract - Specific(50)* resulted in an adjusted p-value of .017, the pairwise comparison of *Specific(5) - Specific(50)* resulted in an adjusted p-value of 0.61. Hence, there is no significant difference in *Takeover Feeling* between *Specific(5)* and *Specific(50)* scenarios, and therefore, no evidence in support of hypothesis H1.3 (see Table 3).

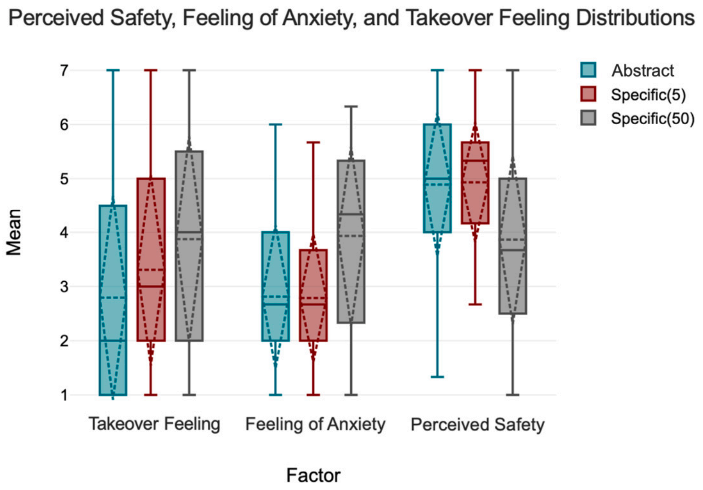


Fig. 6. Perceived safety, feeling of anxiety, and takeover feeling distribution. Perceived safety is highest in the Specific(5) scenario, the feeling of anxiety is highest in the Specific(50), and takeover feeling is lowest in the Abstract scenario.

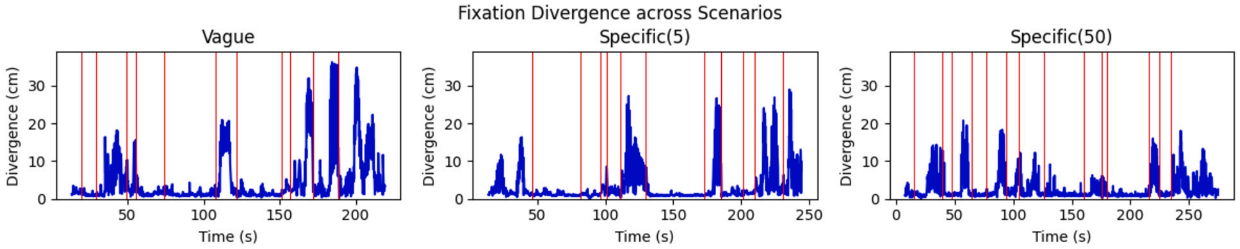


Fig. 7. Fixation divergence across scenarios. While *Specific(5)* had the highest mean fixation divergence, *Specific(50)* had more frequent high fixation divergences. Red vertical bars represent the positions in time where causal explanations were provided.

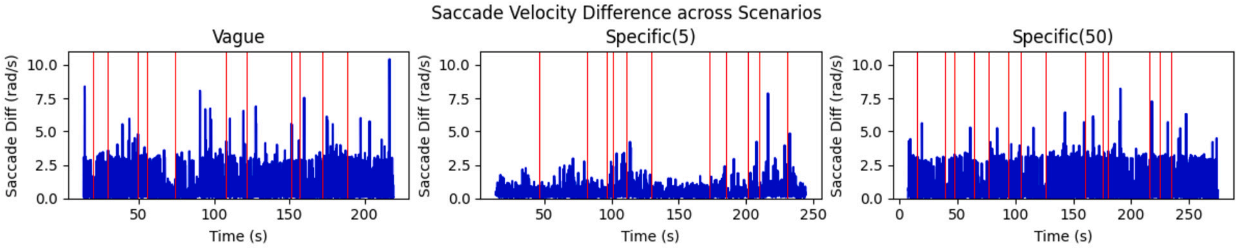


Fig. 8. Saccade velocity difference across scenarios. *Specific(5)* had the lowest mean saccade velocity difference while the *Abstract* scenario had the highest. Red vertical bars represent the positions in time where causal explanations were provided.

5.2. Behavioural cues analysis

H2.1 - visual responses *Visual feedback from passengers correlates with passengers' anxiety*. At this stage, we utilised the reference data from the researcher. We estimated the Euclidean distances between participants' fixation points and the reference fixation points over time for each participant.

Results from Spearman correlation showed that there was no significant association between the *Feeling of Anxiety* and *Fixation Divergence*, $r(115) = -0.07, p = .442$. See the fixation divergence plot in Fig. 7. Results from Spearman correlation showed that there was no significant association between the *Feeling of Anxiety* and *saccade difference*, $r(115) = 0.1, p = .281$. However, there was a significant association between *perceived safety* and *saccade difference*, $r(115) = -0.25, p = .007$., indicating a weak negative correlation between *perceived safety* and *saccade difference*. Hypothesis H2.1, therefore, has no sufficient support. See the saccade difference plot in Fig. 8.

In addition to correlation, we checked for significant differences. There was a significant difference in *Fixation Divergence* between *Abstract* and *Specific(5)* with an adjusted p-value of .028, and between *Specific(5)* and *Specific(50)* with an adjusted p-value $< .001$. See Table 4 for descriptive statistics. Also, there was a significant difference between *Abstract* and *Specific(5)* with respect to *Saccade*

Table 4
Descriptive statistics from the button presses and Visual responses.

	ButtonPress $H(2) = 15.44, \mathbf{p<.001}$			Fixation Divergence $H(2) = 20.67, \mathbf{p<.001}$			Saccade Difference $H(2) = 15.35, \mathbf{p<.001}$		
	Mean	SD	Mean Rank	Mean	SD	Mean Rank	Mean	SD	Mean Rank
Abstract	2.26	3.35	1.72	4.37	1.84	1.95	**1.25**	0.43	**2.42**
Specific(5)	1.64	1.63	1.77	**7.66**	5.21	**2.54**	1.06	0.42	1.54
Specific(50)	**4.9**	4.33	**2.51**	3.2	1.23	1.51	1.17	0.45	2.04

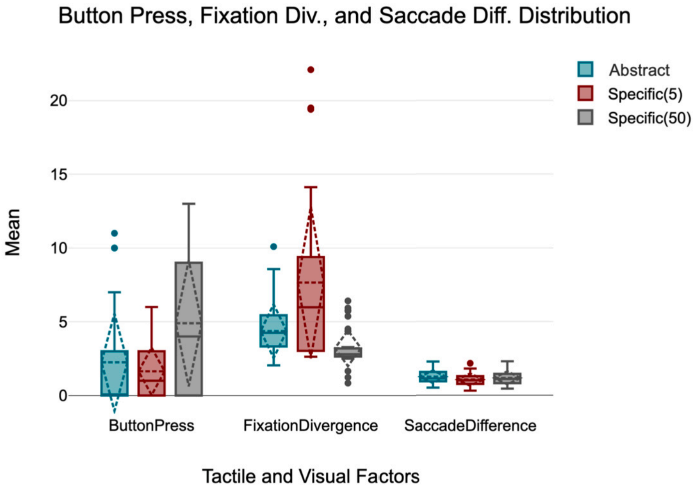


Fig. 9. Button presses, fixation divergence and saccade difference distribution. Button presses are fewer in the *Specific(5)* scenario. Fixation divergence is highest in the *Specific(5)* scenario, and saccade difference is lowest in the *Specific(5)*.

Difference (adjusted p-value of $< .001$). See Fig. 4 for sample scenes from each scenario with the generated explanations. All the participants' gaze points are plotted as heatmaps over the screenshots.

5.2.1. *Button press by passengers*

Participants were asked to press a button on the Logitech wheel when they felt confused, anxious, or unsafe by the explanations or the decision of the AV during the ride. Spearman rank correlation was used as a measure to investigate monotonic associations. There was a weak negative correlation between the variables *Perceived Safety* and *ButtonPress* ($r(115) = -0.31, p = .001$), a weak positive correlation between the *Feeling of Anxiety* and *ButtonPress* ($r(115) = 0.31, p = .001$), and insignificant correlation between the *Feelings to Takeover* and *ButtonPress* ($r(115) = 0.15, p = .099$).

We also checked for statistically significant differences in *Button Presses* across scenarios. There was a significant difference in *ButtonPresses*, $H(2) = 15.44, p < .001$. This was specifically in the pairs: *Abstract - Specific(50)* with adjusted p-value .002, and *Specific(5) - Specific(50)* with adjusted p-value .005. See Fig. 9 for behavioural cues results.

6. Qualitative results: themes and reflections

We obtained qualitative data from the APT questionnaire administered after every scenario. Participants were asked to describe their feelings regarding the explanations they received during the ride. Table 5 and Fig. 10 describe the themes obtained from the inductive thematic analysis of the comments. Themes are broadly categorised based on the participants' feelings, their assessment of the explanations, and the vehicle dynamics.

Perceptual errors in the *Specific(50)* scenario evoked negative emotions of anxiety, feeling to takeover navigation control and distrust. CAND1 expressed a feeling of anxiety: '*The explanations made me feel a bit anxious, it says many things that were not right and misleading. I had the urge to look at the buildings and the environment but could not really do that because I wanted to be sure the vehicle is taking the right decision.*'. CAND39 expressed the urge to takeover navigation control: '*When the explanations are false, e.g. 'a cyclist is crossing my lane', and it is actually a pedestrian, it made me slightly anxious and likely to want to take over. But nevertheless, I felt safe in the vehicle*'. CAND5 expressed distrust in the AV: '*anxious as the vehicle did not correctly understand the environment and the types of vehicles around it, which made me trust its judgement less*'. More participants expressed a feeling of safety in the *Specific(5)* scenario: '*felt safe that the vehicle understood the road and what was going on around us*'. About the same number of participants expressed a decline in their feeling of anxiety in the *Abstract* and *Specific(5)* scenarios. An example is CAND34's comment about the abstract scenario: '*When the explanations provided are more general, e.g. 'vehicle' instead of 'van' and 'road user' instead of 'cyclist', it feels like the vehicle has a better understanding of the surroundings because it gives a correct explanation, so I felt less anxious and unsafe*'. The abstract explanations might have concealed some errors, in turn, reducing the feeling of anxiety.

Table 5
Themes derived from the thematic analysis of the qualitative data from participants. Freq. = Frequency of occurrence, SP = Scenario Percentage.

		Abstract		Specific(5)		Specific(50)	
Category	Theme	Freq.	SP (%)	Freq.	SP (%)	Freq.	SP (%)
Feelings	Anxious	2	5	2	5	8	**21**
	Less Anxious	5	**13**	5	**13**	1	3
	Safe	9	23	12	**31**	7	18
	Unsafe	0	0	1	**3**	1	**3**
	Takeover	2	5	2	5	7	**18**
	Confident	2	5	5	**13**	3	8
	Trust	2	**5**	1	3	2	**5**
	Distrust	1	3	0	0	6	**15**
	Reassuring	5	**13**	2	5	0	0
	Uncomfortable	2	**5**	1	3	0	0
[6pt] Explanations	Good Timing	1	**3**	0	0	0	0
	Bad Timing	7	**18**	1	3	1	3
	Plausible	2	5	10	**26**	1	3
	Implausible	5	13	3	8	25	**64**
	Unintelligible	6	**15**	0	0	0	0
	Repetitive	3	8	4	**10**	2	5
	Vague	5	**13**	0	0	0	0
[6pt] Vehicle Dynamics	Careful Manoeuvre	3	8	2	5	4	**13**
	Aggressive Manoeuvre	1	3	3	8	5	**13**
	Vehicle Feature	0	0	3	**8**	1	3

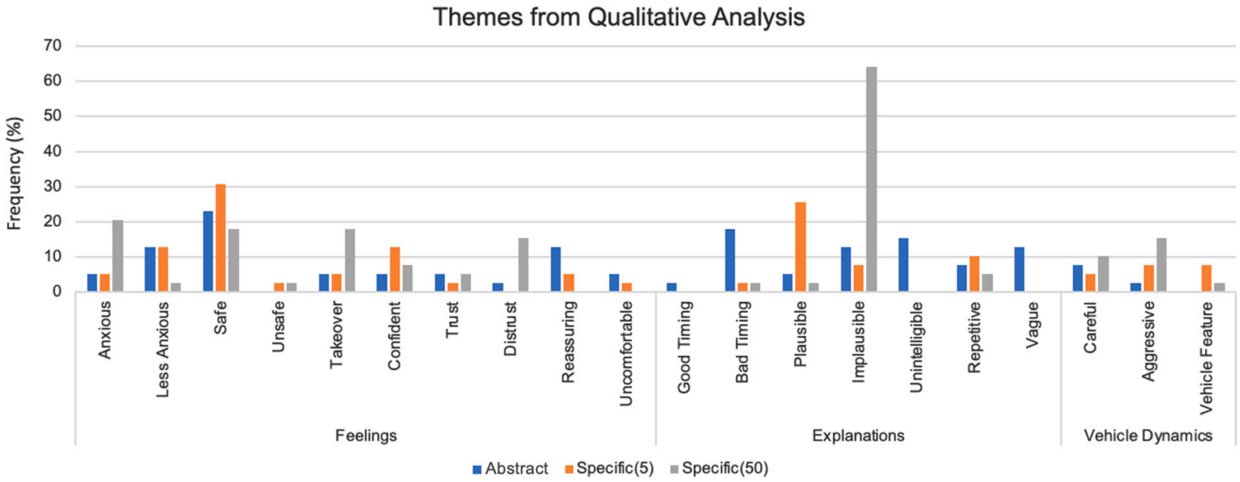


Fig. 10. Themes derived from the thematic analysis of the qualitative data from participants. Frequency is expressed in percentage of the total number of responses in each scenario.

There were specific comments about the explanations across the three scenarios. Many participants thought that the explanations in the *Specific(5)* were plausible in that they sounded correct and aligned with what the participants saw. For example: '*Explanations were clear and made sense. Still don't feel some of the reactions were as quick as I might have made them*'—CAND14. There were a good number of comments around the implausible nature of the explanations in the *Specific(50)* scenario. For example, CAND20 said, '*The vehicle this time had difficulty giving the correct reason for stopping/going. Couldn't tell the difference between a pedestrian and a cyclist sometime or thought that traffic lights were off instead of green. I feel that this time I would have wanted more control over the car, particularly at traffic lights as I could determine better if a traffic light was 'working' or not*'.

A couple of candidates thought that the explanations in the *Abstract* scenarios were either too early or late. For example, '*The explanations should have arrived a bit earlier, like a few meters before the vehicle actually stops so that I will know that it is planning to stop. Also, I would be more comfortable if the explanation 'traffic sign' was 'traffic light is red/green'. when referring to a traffic light.*'—CAND19.

Some interesting comments were made about the vehicle's driving style and its interior. For example, CAND31 made a comment about the careful manoeuvre of the vehicle in the *Specific(50)* scenario: '*I was calm throughout the journey. There was no feeling of anxiety as the vehicle did not speed too much to make me feel that way.*'—CAND31. There was a comment relating to aggressive manoeuvre in the *Abstract* scenario: '*Seemed like oncoming vehicles were going to collide with me. It seems to sometime drive on pavements when negotiating corners.*'—CAND35. The rotating steering wheel of the vehicle made some of the participants uncomfortable: 'The steering wheel moving abruptly startled me sometimes.'—CAND21 (*Specific(5)* scenario). Some participants liked the vehicle indicators and the sound they made when indicating the next directions. '*The indicator sound was nice to hear. [...]*'—CAND6 (*Specific(50)* scenario).

7. General discussion

We examined the effects of explanation specificity (*abstract* and *specific*) in AD, while accounting for varying degrees of perception system detection errors (*low* and *high*). We focused on how this setup would impact passengers' perceived safety and related factors, such as the feeling of anxiety and the thought to takeover control. Our results not only corroborate but also extend previous findings in the field, among others, demonstrating that while intelligible explanations generally create positive experiences for AV users (Omeiza et al., 2021b; Ha et al., 2020; Schneider et al., 2021; Faas S et al., 2021), this effect is predominantly observed when the AV's perception system detection errors are low.

7.1. Psychological effects

Hypothesis 1.1 - low transparency yields higher perception of safety Participants expressed a greater but statistically insignificant sense of safety in the *Specific(5)* scenario compared to *Abstract* scenario. We also see from the qualitative result on Table 5 that the difference in the number of responses attributed to unsafe feelings between *Abstract* and *Specific(5)* is only 1. Hence, abstract explanations (low transparency) did not lead to higher perception of safety.

On a different note, we observed a statistically significant difference in perceived safety between *Specific(5)* and *Specific(50)*, see Table 3 and Fig. 6. AVs characterised by high transparency and high perception system detection errors generally elicited lower perceptions of safety among passengers. This result aligns with previous work suggesting that transparency can sometimes amplify the negative effects of errors on trust in automated systems (De Visser et al., 2018).

From our qualitative results, overly detailed explanations were perceived as verbose and repetitive by some participants in the specific scenarios (both *Specific(5)* and *Specific(50)*). Therefore, achieving a balance between explanation specificity (or transparency in general) and the cognitive load imposed on passengers is crucial (Poursabzi-Sangdeh et al., 2021). Rather than being vague, as in the abstract explanations, this might mean further simplification of the explanations while still retaining essential information.

In our study, the combination of detailed explanations and frequent errors may have heightened passengers' awareness of the system's limitations, potentially exacerbating their safety concerns. Nonetheless, qualitative responses revealed that a minority of participants reported positive sentiments despite these errors. They praised the vehicle for its ability to detect obstacles and respond appropriately, suggesting that for these individuals, the AV's correct decision-making outweighed concerns about the specific type of obstacle encountered.

The diverse reactions we observed highlight the complex nature of human-AV interaction and underscore the need for personalised approaches to transparency and explanation design. Adaptive explanation strategies that consider individual differences in information processing and risk perception may be necessary to optimise user experience and safety perceptions in autonomous vehicles.

Hypothesis 1.2 - feeling of anxiety increases with increasing perception system detection errors Drivers' anxiety has been noted to increase when utilising AVs (Koo et al., 2016). In our study, we anticipated that passengers' feelings of anxiety would correspondingly increase with higher levels of perception system detection errors in an AV. This hypothesis was supported by a significant difference in anxiety levels observed between the *Specific(5)* and *Specific(50)* scenarios. In the context of AVs, perception system detection errors represent a critical factor that can directly impact passenger safety and, consequently, their psychological state. As the frequency of errors increases, passengers may experience a heightened sense of uncertainty and loss of control. A related study (Jansen et al., 2024) noted this trend as participants were sceptical of the detection and prediction capabilities of AVs for challenging scenes after observing the performances of the underlying algorithms in an experimental setting. These increased errors might be a contributor to the reported anxious feelings.

Given that we had higher values for perceived safety in *Specific(5)* scenario (Table 3), we expected the lowest levels of anxiety in this scenario, assuming a relationship between perceived safety and anxiety, as suggested by prior research (Dillen et al., 2020). This inverse relationship between perceived safety and anxiety aligns with the broader psychological concept of risk perception and its impact on emotional states (Stapel et al., 2022; Slovic, 2010). Our findings extend beyond the mere confirmation of this relationship, highlighting the nuanced interplay between system transparency, error frequency, and the feeling of anxiety. The significant difference in anxiety levels between the *Specific(5)* and *Specific(50)* scenarios suggests that there may be a threshold of error frequency beyond which anxiety levels sharply increase. This notion is supported by research in risk perception, which indicates that individuals often have a non-linear response to increasing risk levels (Fischhoff et al., 1993). Investigating this threshold is an interesting topic for future research.

While Dillen et al. (2020), primarily examined how AV driving styles affect driver anxiety and comfort, they noted that certain in-vehicle features, such as a rotating steering wheel, could also influence feelings of anxiety among participants. This experience was reported by participants (e.g., CAND21) in our study. This observation underscores the multifaceted nature of anxiety in AV contexts, where both system performance and physical design elements play crucial roles.

Furthermore, our results suggest that the relationship between perception system detection errors and anxiety may be moderated by the level of explanation specificity. In scenarios with high transparency (i.e., specific explanations), the impact of errors on anxiety might be more pronounced, as passengers are more acutely aware of the system's limitations. On the other hand, uncertainty about the workings of algorithms and perceived lack of control has been noted to cause what researchers term 'algorithmic anxiety' (Eiband et al., 2019b). This is also observed from our result in Fig. 6 as the feeling of Anxiety is higher in the *Abstract* scenario compared to *Specific(5)*.

It is worth noting that individual differences in technology acceptance and risk tolerance may also influence the anxiety response to perception system detection errors (Choi and Ji, 2015). Future research could explore these individual factors to develop a more comprehensive model of anxiety in AV contexts.

Hypothesis 1.3 - takeover feeling increases with the increase in perception system error Contrary to our hypothesis, the data did not support an increase in takeover feeling with increased perception system detection errors. While we observed a significant difference between the *Abstract* and *Specific(50)* scenarios, the lack of significant difference between *Specific(5)* and *Specific(50)* scenarios suggests a complex relationship between system errors and user responses. This finding indicates that mere disclosure of perception system detection errors does not necessarily escalate passengers' desire for control, pointing to a possible threshold effect in how passengers perceive and respond to AV errors. These results challenge the notion that awareness of system imperfections automatically erodes trust in automation (Lee & See, 2004), suggesting instead that passengers may be more tolerant of disclosed errors than previously thought, especially with regard to seizing control in autonomous driving.

Our empirical findings contrast with the conceptual analysis presented by Terken and Pfleging (2020), who advocated for shared control between vehicle and user. Our results suggest that users might be more comfortable with full automation than theoretical analyses have predicted, even when aware of system limitations. It is crucial to reconcile these findings with our previous results showing increased anxiety levels as error rates increased. This apparent contradiction suggests a complex interplay between emotional responses and behavioural intentions in AV contexts. While passengers may experience heightened anxiety feelings when aware of higher error rates, this emotional response does not necessarily translate into a stronger desire to take control of the vehicle. This disconnect between anxiety and takeover feeling could be attributed to factors such as perception of one's own ability to manage the situation, the cognitive dissonance between acknowledging risks and maintaining comfort with automation (Aronson, 1969), or trust in the system's overall capability.

It is important to note, on the other hand, that our qualitative results (Table 5) indicate a higher number of comments relating to takeover intents in *Specific (50)* compared to *Specific (5)*. It might be the case that participants might have struggled to calibrate their thoughts or feelings on a 5-point Likert scale. It might also be the case that this difference in the frequency of the said comments is statistically insignificant. This would be investigated in future work.

7.2. Behavioural cues

Hypothesis 2.1 - visual signal correlates with anxiety Our study's findings challenge the established link between anxiety and visual distraction proposed by Hepsomali et al. (2017). We found no significant correlation between fixation point divergences and anxiety levels across the different scenarios. This surprising outcome might be attributed to the varied individual priorities in visual attention. For instance, while some participants might have focused on the cityscape, others may have concentrated on areas highlighted by the explanations. The *Specific(50)* scenario exhibited higher divergences in fixation points compared to the *Specific(5)* and *Abstract* scenarios, possibly due to misclassifications in explanations directing attention to incorrect elements. However, the *Specific(5)* and *Abstract* scenarios showed similar fixation effects, suggesting that explanations might have been more effective in these scenarios than in the *Specific(50)* scenario. Contrary to Dillen et al., Dillen et al. (2020)'s findings on the relationship between eye movement entropy and anxiety, our study did not reveal a significant correlation between saccade differences and anxiety levels. Interestingly, the *Specific(5)* scenario demonstrated the lowest saccade difference, potentially indicating reduced distraction or confusion. This assumption is based on the understanding that saccade velocity reflects the speed of gaze shifts between fixation points. In contrast, the *Abstract* scenario exhibited the highest saccade difference, which could be interpreted as more active visual searching due to the non-specific nature of the explanations failing to effectively guide participants' gaze.

These findings amplify the existing complexities in the relationship between visual behaviour, anxiety, and the nature of explanations in automated driving scenarios. Further research to better understand the interplay between cognitive states, visual attention, and the effectiveness of different types of explanations in automated driving contexts is in order. Additionally, investigating the potential impact of individual differences in visual processing and attention allocation could provide valuable insights into designing more effective and personalised explanation systems for automated vehicles.

7.3. Practical implications

Our findings challenge the initial assumption that passengers may not desire specific explanations detailing error information from automated vehicles (AVs). Contrary to this presupposition, the study reveals a preference among passengers for specific explanations, particularly when the AV's perception system demonstrates near-perfect accuracy. This insight has significant implications for both manufacturers and regulators in the AV industry. The observed inverse relationship between perception system detection errors and passenger anxiety underscores the critical need for highly transparent AVs with exceptional perception and decision-making capabilities. This finding aligns with the broader literature on trust in automation, which emphasises the importance of system reliability and transparency in fostering user acceptance (Lee & See, 2004). Manufacturers should prioritise the development of robust perception systems that minimise errors, while simultaneously implementing transparent communication mechanisms to convey system status and decision-making processes to passengers. Regulators, in turn, should consider establishing or strengthening existing standards for AV perception accuracy and mandating clear and usable interfaces for conveying this information to passengers.

While our study did not observe direct consequences of misclassification errors on AV actions, it is crucial to recognise the potential implications of such errors in more complex, real-world scenarios. The accurate classification of obstacle types is paramount in determining appropriate vehicle responses, particularly when dealing with dynamic obstacles possessing varied manoeuvrability characteristics. This aligns with research efforts on situation awareness in automated systems, which emphasises the importance of accurate environmental perception for effective decision-making (Endsley, 1995). In intricate traffic scenarios, even minor detection inaccuracies could lead to sub-optimal navigation decisions, potentially compromising safety. Therefore, AV developers must strive for highly accurate environmental estimations to ensure appropriate responses to diverse obstacle types, such as bicycles or motorcycles, with very similar attributes but differing capabilities.

The study's findings highlight the link between transparency and accuracy in AV systems. This relationship echoes the concept of 'calibrated trust' in automation, where user trust is appropriately aligned with system capabilities (Lee & See, 2004). To foster this calibrated trust, AV interfaces should not only provide accurate information but also communicate the system's confidence levels and potential limitations. This approach can help manage passenger expectations and maintain appropriate levels of situational awareness.

Transparency should aim to empower the relevant stakeholders with sufficient information to make informed choices and avoid potential misuse or misunderstanding of the system. In addition to enhancing AVs with explanations for situational awareness, other critical information, such as safety notes and reports about the various components and functional modes of the AV, should be made available, and principles of honesty, user autonomy, and informed consent should guide such.

Finally, these findings have implications for AV design, regulation, and user experience. By prioritising transparency and accuracy and adopting a holistic approach to user research, the AV industry can work towards creating systems that meet technical performance standards and passengers' needs while factoring in likely psychological effects.

8. Limitations and future work

Our work has some limitations: first, as mentioned in the previous section, transparency should aim to empower stakeholders with sufficient information to make informed choices. However, this work did not provide the choice for participants to decide the amount of details they would like in an explanation during a drive. Future work would incorporate personalisation capabilities to meet the information needs of the different stakeholder types.

Second, although the VR environment provided an immersive and near-realistic driving experience, the experience still falls short in comparison with actual real-life driving and the use of vehicle simulation with motion feedback.

Third, we note that there are various factors, including culture and gender, that could influence driving styles. In the future, we will ensure better representation and gender balance while recruiting participants.

Fourth, while this study focuses on the impact of explanation specificity—building on prior research, discussed in Section 2.2, that highlights that explanations are valuable in self-driving—it would be interesting to compare all the scenarios examined here to a baseline case without explanations. We acknowledge the absence of this baseline case as a limitation of the current work.

Fifth, while visual feedback from experimental studies provides valuable insights into participants' psychological and behavioural responses, it is useful to simultaneously implement multiple methodologies (e.g., mixed methods) for more accurate conclusions. Complementing observational data with other measurement techniques, such as standardised surveys, physiological measurements, or qualitative interviews, can provide a more comprehensive understanding of user experiences and perceptions. This multi-faceted approach allows for triangulation of data and more confident interpretations of complex human-machine interactions.

Lastly, future research should explore the long-term effects of exposure to AV explanations on passenger trust, anxiety, and overall acceptance. Longitudinal studies could reveal how user preferences and responses evolve over time, informing the design of adaptive explanation systems that cater to changing user needs and expectations. Additionally, investigating the impact of cultural differences and individual variability in risk perception on AV explanation preferences could yield insights for developing globally applicable yet locally tailored AV communication strategies.

9. Conclusion

In this study, we conducted a rigorous within-subject laboratory investigation ($N = 39$) utilising an immersive driving simulator to examine passengers' reactions to explanation specificity levels with varying perception system detection errors in automated driving (AD). We observed a significant increase in anxiety levels when fine-grained, intelligible explanations exposed perception system detection errors. However, these errors did not have any significant influence on their feeling to takeover control of the AV. We also observed that abstract explanations (despite the potential to conceal errors) did not lead to an increased perception of safety compared to specific explanations that exposed minimal amount of detection errors. Rather, specific explanations with minimal errors yielded greater but statistically non-significant perception of safety. We referred to this phenomenon as the *transparency paradox* in the context of this paper. The lack of substantial evidence to clearly unravel the connection between transparency and perceived safety when perception error is minimal (as in a realistic scenario) necessitates further studies. Our work contributes to the growing body of literature on trust calibration and transparency in automated vehicles, emphasising the need for adaptive, context-sensitive explanation systems that offer the stakeholders' desired level of transparency and comfort, ultimately fostering greater public acceptance and trust in autonomous transportation systems.

CRediT authorship contribution statement

Daniel Omeiza: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Raunak Bhattacharyya:** Writing – review & editing, Writing – original draft, Data cura-

tion. **Marina Jirotka:** Supervision, Resources, Funding acquisition. **Nick Hawes:** Supervision, Resources, Funding acquisition. **Lars Kunze:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Perplexity AI and ChatGPT to search for related papers, improve grammar, and shorten certain texts to meet requirements. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Acknowledgements

This work was supported by the EPSRC RAILS project (grant reference: EP/W011344/1) and the EPSRC project RoboTIPS (grant reference: EP/S005099/1).

Appendix A. Online survey questionnaire

Note that the same questionnaire was administered in all scenarios.

Data availability

Data will be made available on request.

References

Call For Participants, https://www.callforparticipants.com/. (Accessed 24 July 2022).

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Anjomshoae, S., Omeiza, D., & Jiang, L. (2021). Context-based image explanations for deep neural networks. *Image and Vision Computing*, *116*, Article 104310.

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology. Elsevier.: Vol. 4* (pp. 1–34).

Bin Issa, R., Das, M., Rahman, M. S., Barua, M., Rhaman, M. K., Ripon, K. S. N., & Alam, M. G. R. (2021). Double deep Q-learning and faster R-Cnn-based autonomous vehicle navigation and obstacle avoidance in dynamic environment. *Sensors*, *21*, 1468.

Board, N. T. S. (2018). Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator mountain view. California. URL, https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf. (Accessed 30 October 2020).

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines*, *32*, 563–584.

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020). The emerging landscape of explainable automated planning & decision making. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20, international joint conferences on artificial intelligence organization* (pp. 4803–4811).

Chen, N. T., Clarke, P. J., Watson, T. L., Macleod, C., & Guastella, A. J. (2014). Biased saccadic responses to emotional stimuli in anxiety: An antisaccade study. *PLoS ONE*, *9*, Article e86474.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*, 692–702.

Colley, M., Bräuner, C., Lanzer, M., Walch, M., Baumann, M., & Rukzio, E. (2020). Effect of visualization of pedestrian intention recognition on trust and cognitive load. In *12th international conference on automotive user interfaces and interactive vehicular applications* (pp. 181–191).

Colley, M., Eder, B., Rixen, J. O., & Rukzio, E. (2021). Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–11).

Colley, M., Rädler, M., Glimmann, J., & Rukzio, E. (2022). Effects of scene detection, scene prediction, and maneuver planning visualizations on trust, situation awareness, and cognitive load in highly automated vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*, 1–21.

Davidson, M. M., Butchko, M. S., Robbins, K., Sherd, L. W., & Gervais, S. J. (2016). The mediating role of perceived safety on street harassment and anxiety. *Psychology of Violence*, *6*, 553.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation'to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*, 1409–1427.

Dillen, N., Ilievski, M., Law, E., Nacke, L. E., Czarnecki, K., & Schneider, O. (2020). Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13).

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO)*.

Dosovitskiy, A., Ros, G., & Codevilla, F., Lopez, A., Koltun, V. CARLA: An open urban driving simulator. In *Conference on robot learning* (2017).

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019a). The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1–6).

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019b). When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 96–106).

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64.

Endsley, M. R. (2000). Situation models: An avenue to the modeling of mental models. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 61–64). Los Angeles, CA: SAGE Publications Sage CA.

Fischhoff, B., Bostrom, A., Quadrel, M. J., et al. (1993). Risk perception and communication. *Annual Review of Public Health*, *14*, 183–203.

Gibaldi, A., & Sabatini, S. P. (2021). The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis. *Behavior Research Methods*, *53*, 167–187.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*, 1–42.

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*.

Ha, T., Kim, S., Seo, D., & Lee, S. (2020). Effects of explanation types and perceived risk on trust in autonomous vehicles. *Transportation Research. Part F, Traffic Psychology and Behaviour*, *73*, 271–280.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*, 555.

Hepsomali, P., Hadwin, J. A., Liversedge, S. P., & Garner, M. (2017). Pupillometric and saccadic measures of affective and executive processing in anxiety. *Biological Psychology*, *127*, 173–179.

Hewitt, C., Politis, I., Amanatidis, T., & Sarkar, A. (2019). Assessing public perception of self-driving cars: The autonomous vehicle acceptance model. In *Proceedings of the 24th international conference on intelligent user interfaces*.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, *32*, 68–73.

Information Commissioner's Office (2021). What goes into an explanation?. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation/. (Accessed July 2021).

Jansen, P., Colley, M., Pfeifer, T., & Rukzio, E. (2024). Visualizing imperfect situation detection and prediction in automated vehicles: Understanding users' perceptions via user-chosen scenarios. *Transportation Research. Part F, Traffic Psychology and Behaviour*, *104*, 88–108.

Kaufman, R., Costa, J., & Kimani, E. (2024). Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust. *Scientific Reports*, *14*, Article 13061.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research. Part C, Emerging Technologies*, *96*, 290–303.

Kment, B. (2006). Counterfactuals and explanation. *Mind*, *115*, 261–310.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). 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)*, *9*, 269–275.

Koo, J., Shin, D., Steinert, M., & Leifer, L. (2016). Understanding driver responses to voice alerts of autonomous car operations. *International Journal of Vehicle Design*, *70*, 377–392.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019a). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*, 345–360.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019b). Conveying uncertainties using peripheral awareness displays in the context of automated driving. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications* (pp. 329–341).

Lavrinc, D. (2016). This is how bad self-driving cars suck in rain. https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/. (Accessed July 2020).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.

Liu, H., Hirayama, T., & Watanabe, M. (2021). Importance of instruction for pedestrian-automated driving vehicle interaction with an external human machine interface: Effects on pedestrians' situation awareness, trust, perceived risks and decision making. In *2021 IEEE intelligent vehicles symposium (IV)* (pp. 748–754).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*.

M. Faas, S., Kraus, J., Schoenhals, A., & Baumann, M. (2021). Calibrating pedestrians' trust in automated vehicles: Does an intent display in an external HMI support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI conference on human factors in computing systems*.

Magnaguagno, M. C., Fraga Pereira, R., Móre, M. D., & Meneguzzi, F. R. (2017). Web planner: A tool to develop classical planning domains and visualize heuristic state-space search. In *2017 workshop on user interfaces and scheduling and planning (UISP@ ICAPS)*.

McFarland, M. (2016). Who's responsible when an autonomous car crashes? https://money.cnn.com/2016/07/07/technology/tesla-liabilityrisk/index.html. (Accessed 24 July 2020).

Michael, M., & Schlipsing, M. (2015). Extending traffic light recognition: Efficient classification of phase and pictogram. In *2015 international joint conference on neural networks (IJCNN)*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*.

Nees, M. A., Helbein, B., & Porter, A. (2016). Speech auditory alerts promote memory for alerted events in a video-simulated self-driving car ride. *Human Factors*, *58*, 416–426.

Omeiza, D., Anjomshoae, S., Webb, H., Jirotka, M., & Kunze, L. (2022a). From spoken thoughts to automated driving commentary: Predicting and explaining intelligent vehicles' actions. In *2022 IEEE intelligent vehicles symposium (IV)*.

Omeiza, D., Kollnig, K., Webb, H., Jirotka, M., & Kunze, L. (2021a). Why not explain? Effects of explanations on human perceptions of autonomous driving. In *IEEE international conference on advanced robotics and its social impacts*.

Omeiza, D., Web, H., Jirotka, M., & Kunze, L. (2021b). Towards accountability: Providing intelligible explanations in autonomous driving. In *2021 IEEE intelligent vehicles symposium (IV)*.

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2022b). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*, 10142–10162. https://doi.org/10.1109/TITS.2021.3122865.

Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013). Interpreting random forest models using a feature contribution method. In *2013 IEEE 14th international conference on information reuse & integration (IRI)* (pp. 112–119). IEEE.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the CHI conference on human factors in computing systems*.

Quansah, F., Hagan Jr, J. E., Sambah, F., Frimpong, J. B., Ankomah, F., Srem-Sai, M., Seibu, M., Abieraba, R. S. K., & Schack, T. (2022). Perceived safety of learning environment and associated anxiety factors during COVID-19 in Ghana: Evidence from physical education practical-oriented program. *European Journal of Investigation in Health, Psychology and Education*, *12*, 28–41.

Raab, E. L. (1985). Normal saccadic velocities. *Journal of Pediatric Ophthalmology & Strabismus*, *22*, 20–22.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

SAE, I. (2021). SAE levels of driving automation™ refined for clarity and international audience.

Schneider, T., Hois, J., Rosenstein, A., Ghellal, S., Theofanou-Fülbier, D., & Gerlicher, A. R. (2021). Explain yourself! Transparency for positive ux in autonomous driving. In *Proceedings of the CHI conference on human factors in computing systems*.

Silvera, G., Biswas, A., & Admoni, H. (2022). DReyeVR: Democratizing virtual reality driving simulation for behavioural & interaction research. In *ACM/IEEE human robot interaction conference*.

Slovic, P. (2010). The feeling of risk. Earthscan.

Stanton, N. A., Salmon, P. M., Walker, G. H., & Stanton, M. (2019). Models and methods for collision analysis: A comparison study based on the uber collision with a pedestrian. *Safety Science*, *120*, 117–128.

Stapel, J., Gentner, A., & Happee, R. (2022). On-road trust and perceived risk in level 2 automation. *Transportation Research. Part F, Traffic Psychology and Behaviour*, *89*, 355–370.

Terken, J., & Pfleging, B. (2020). Toward shared control between automated vehicles and users. *Automotive Innovation*, *3*, 53–61.

Tilley, A. (2016). Google's self-driving car caused its first crash. https://www.forbes.com/sites/aarontilley/2016/02/29/googles-self-driving-car-caused-its-first-accident/?sh=5ae097b0538d. (Accessed 21 June 2021).

Voigt, P., & dem Bussche, A. (2017). *The EU general data protection regulation (GDPR). a practical guide* (1st ed). Cham: Springer International Publishing.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the CHI conference on human factors in computing systems*.

Wilson, S. J., Glue, P., Ball, D., & Nutt, D. J. (1993). Saccadic eye movement parameters in normal subjects. *Electroencephalography and Clinical Neurophysiology*, *86*, 69–74.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *IEEE conference on computational intelligence*. and Games (CIG).