Providing and Assessing Intelligible Explanations in Autonomous Driving

Daniel Omeiza

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

To my lovely wife Kala Omeiza, parents,
and ultimately to God through whose grace I am able to complete this thesis.

Acknowledgements

About this time in Autumn 2019, I was walking past the different stalls in Examination Schools during the freshers' fair, grabbing all the pens and perks, signing up to all the university clubs and societies, and confidently describing my intended research to everyone like I had everything figured out. Many weeks down the line, I had presented research questions with different 'flavours' to my supervisors and eventually settled on the problem of explainability in autonomous driving which finally gave birth to this thesis. The credit goes to my patient and very supportive supervisors. Indeed, this journey would not have been possible without the amazing support system around me. First, I want to appreciate God who has given me life and the strength to conduct my research. My sincere appreciation goes to my supervisors, Dr. Lars Kunze and Prof. Marina Jirotka who have been so supportive all through my research endeavours at Oxford. Their thoughts and ideas have been invaluable; they have greatly helped to develop my research career. Our conversations were always thought-provoking, leaving me with high motivation and enthusiasm to do great research.

I give special thanks to Helena Webb for her kind support and guidance even after leaving Oxford to join the University of Nottingham; I thank Helena for sparing some moments to chat about my research and life in general whenever we met along Cowley Road. I appreciate Jun Zhao, Nick Hawes, and Reuben Binns for providing feedback on my research reports. Special thanks to Konrad Kollnig and Siddhartha Datta first for their friendship, constructive criticisms, encouragement and joint efforts to work through challenges along the way. I also want to appreciate my colleague Carolyn Ten Holter who read some parts of this thesis and provided constructive feedback. She has also been so helpful in sharing useful information for my project since we started together in 2019. Thanks to Raunak Bhattacharyya for reviewing some chapters of this thesis and for also participating in my lab experiment.

It was always exciting catching up with friends in the Responsible Research and Innovation group; the game nights, lunch moments etc. were memorable. Thanks to Ross Gales, Paula Fiddi, Towera Moyo, Lize Alberts, and Pericle Salvini. Special thanks to the members of the Cognitive Robotics Group—Ricardo Cannizzaro, Rhys Howard, Efimia Panagiotaki, and Divya Thuremella—for the amazing moments

together and their participation in my lab study. I also want to appreciate Jonathan Attias for his support with setting up demos.

Sincerest gratitude to my wife Kala Omeiza who stood by me when I had to work into the nights at Linacre college and in the ORI to meet paper deadlines. Her words and actions were encouraging, and her patience in listening to my research woes is unmatched. I especially appreciate my parents and sisters for the calls and consistent follow-ups regarding my well-being in the UK. Indeed, my mother, being a computer scientist, has provided all that I needed to become a seasoned computer scientist. Lastly, I appreciate the members of the Deeper Life Bible Church Oxford for their encouragement and their participation in my lab study.

Funding for this doctoral work was made possible by the Department of Computer Science through Prof. Marina Jirotka's UK Engineering and Physical Sciences Research Council (EPSRC) Established Career Fellowship on RoboTIPS: Developing Responsible Robots for the Digital Economy, under Grant EP/S005099/1.

Abstract

Intelligent vehicles with automated driving functionalities provide many benefits, but also instigate serious concerns around human safety and trust. While the automotive industry has devoted enormous resources to realising vehicle autonomy, there exist uncertainties as to whether the technology would be widely adopted by society. Autonomous vehicles (AVs) are complex systems, and in challenging driving scenarios, they are likely to make decisions that could be confusing to end-users. As a way to bridge the gap between this technology and end-users, the provision of explanations is generally being put forward. While explanations are considered to be helpful, this thesis argues that explanations must also be intelligible (as obligated by the GDPR Article 12) to the intended stakeholders, and should make causal attributions in order to foster confidence and trust in end-users. Moreover, the methods for generating these explanations should be transparent for easy audit. To substantiate this argument, the thesis proceeds in four steps: First, we adopted a mixed method approach (in a user study $N = 101$) to elicit passengers' requirements for effective explainability in diverse autonomous driving scenarios. Second, we explored different representations, data structures and driving data annotation schemes to facilitate intelligible explanation generation and general explainability research in autonomous driving. Third, we developed transparent algorithms for posthoc explanation generation. These algorithms were tested within a collision risk assessment case study and an AV navigation case study, using the Lyft Level5 dataset and our new SAX dataset—a dataset that we have introduced for AV explainability research. Fourth, we deployed these algorithms in an immersive physical simulation environment and assessed (in a lab study $N = 39$) the impact of the generated explanations on passengers' perceived safety while varying the prediction accuracy of an AV's perception system and the specificity of the explanations. The thesis concludes by providing recommendations needed for the realisation of more effective explainable autonomous driving, and provides a future research agenda.

Contents

List of Figures

List of Tables

List of Abbreviations

AI . . . . . Artificial Intelligence

ADAS . . . . . Advance Driving Assistant System

AV . . . . . Autonomous Vehicle

CAV . . . . . Connected Autonomous Vehicles

IV . . . . . Intelligent Vehicle

XAI . . . . . Explainable AI

HMI . . . . . Human-Machine Interface

eHMI . . . . . External Human-Machine Interface

VR . . . . . Virtual Reality

SAX . . . . . Sense-Assess-eXplain

CI . . . . . Contextual Importance

DT . . . . . Decision Tree

RF . . . . . Random Forest

CAN . . . . . Controller Area Network

GDPR . . . . . General Data Protection Right

ICO . . . . . Information Commissioner's Office

CAM . . . . . Class Activation Map

Grad-CAM . . Gradient Class Activation Map

EDR . . . . . Event Data Recorder

ITS . . . . . Intelligent Transport Systems

SAE . . . . . Society of Automobile Engineers

PRM . . . . . Pedestrian with Reduced Mobility

TTC . . . . . Time to Collision

CNN . . . . . Convolutional Neural Network

IMU . . . . . Inertial Measurement Unit

GNSS . . . . . Global Navigation Satellite System

GPS . . . . . Global Positioning System

FE1 . . . . . Factual Explanation Algorithm

FE2 . . . . . Improved Factual Explanation Algorithm

CFE1 . . . . . Counterfactual Explanation Algorithm

CFE2 . . . . . Improved Counterfactual Explanation Algorithm

1 Introduction

The automotive industry has witnessed an increasing level of development in the past years; from manually operated vehicles to vehicles with a high level of automation. Despite the technological advancements, accidents caused by nascent technologies such as autonomous vehicles (AVs) (Lavrinc, 2016; McFarland, 2016; NTBS, 2018; Stanton et al., 2019; Tilley, 2016) continue to hamper public trust (Yurtsever et al., 2020a). With the hope to deploy intelligent vehicles (IV), and in particular, autonomous vehicles (AVs) on a commercial scale, the acceptance and willingness of society to use AVs become paramount and may largely depend on AVs' degree of transparency, trustworthiness, safety, and compliance with regulations. Moreover, in challenging driving scenarios, AVs are likely to make decisions that are confusing to end-users; these decisions are consequential in critical situations. Hence, effective ways to build confidence, trust and provide assurance of safety to end-users are desired. One of such ways is the provision of explanations (Ha et al., 2020; Koo et al., 2015). For explanations to be helpful in achieving the aforementioned goals, we argue that they should be intelligible (as obligated by the GDPR Article 12 (Voigt & Von dem Bussche, 2017) to the intended stakeholder. Indeed, explainability should be regarded as a critical requirement for AVs. AVs should be able to explain what they have 'seen', done, and might do in environments in which they operate. Besides benefits to end-users, explanations can also benefit other stakeholders—such

as incident investigators—when the explanations are comprehensive, and their correctness and faithfulness can be guaranteed. Consider the Molly problem (ITU, 2020) described by the International Telecommunication Union (ITU) in relation to AVs. A young girl called Molly was crossing the road alone and was hit by an unoccupied self-driving vehicle. There were no eyewitnesses. Comprehensive and faithful post-hoc explanations containing the vehicle's observations, the road rules, and the traffic signs the vehicle acted on will serve as clues to the causes of the accident. These clues will inform the accident investigation process. Moreover, system auditors can also benefit from an easier auditing process with highly detailed explanations provided over time. However, in this thesis, we focused on investigating more high level and intelligible explanations as the target recipients are passengers.

As highly automated vehicles make high-stake decisions that can significantly affect passengers, we expect them to explain or justify their decisions to reassure safety and assist passengers in appropriately calibrating their trust.

1.1 Fundamental Issues

1.1.1 Limited Human Centric XAI Research in Autonomous Driving

Explainable artificial intelligence (XAI) research is seen to be on the rise. Despite this growth, the applications of XAI are limited in some critical domains, especially, in autonomous driving. A large body of literature in explainable AI focuses on explaining a single artificial neural network model (Chattopadhay et al., 2018; Omeiza et al., 2019; Selvaraju et al., 2017) while only a handful focuses on explaining a goal-based system, such as autonomous vehicles which possess unique architecture and various interacting sub-systems.

In practice, human-machine interfaces in automated vehicles sometimes provide visualisations of the perception systems' view of the world and planned trajectory in ways relatable to the in-vehicle participants (Gillmore & Tenhundfeld, 2020). However, causal explanations that respond to implicit questions of the form of 'Why', 'Why-Not', 'What-If' are not present. In the academic literature, attribution

(or influence score) techniques that employ heatmaps (e.g., saliency and attention heatmaps) have been proposed as a way to respond to implicit 'Why questions' in autonomous driving (Bojarski et al., 2018; J. Kim et al., 2018; Xu et al., 2020). Attribution or influence score approaches explain the decision of a black-box model by providing continuous values whose magnitude indicates the importance of each input feature for a given prediction (J. Zhou et al., 2021). *Saliency* and *attention heatmaps* help to provide a visual representation of these values over the input space. These works suffer from poor *intelligibility*, especially when lay users (e.g., passengers) are involved.

1.1.2 Intelligibility

Article 12 of the GDPR (Voigt & Von dem Bussche, 2017) demands that information be provided to data subjects in an intelligible construct. The term *intelligibility* is used to describe how easily an explanation could be understood or comprehended (Lim et al., 2009). While many existing AI systems which are considered interpretable could be understood by experts, a thorough investigation of their properties shows no indication of intelligibility when lay users are involved (Chakraborti, Kambhampati, et al., 2017; Sreedharan et al., 2017). Further, many explainable AI algorithm design processes are not informed by users' needs. As an example, Chakraborti, Kambhampati, et al. (2017) proposed an algorithm for explaining the plans of an autonomous robot. The explanations generated by this algorithm are not communicated in natural language; thus, they are not easy to comprehend by lay users. This is a serious concern, especially for social robots and autonomous vehicles (AVs) where the party that requires explanations to enhance understanding or build trust is usually not an AI expert. A visual explanations (e.g., saliency and attention heatmaps) which show where the deep driving model is attending would hardly pass a clear message that can draw a passenger or driver's attention for an immediate response. Moreover, saliency explanation methods are noted to create spurious heatmaps, with high entropy. This is evidenced in the seminal work of Adebayo et al. (2018) on sanity checks for saliency methods.

In summary, the limitations in the application of explainable AI in the autonomous driving domain where agents are goal-based with complex architectures is a key motivation for this research. More importantly, the intelligibility of explanations is critical in autonomous vehicles. Hence, explainable AI algorithms should be designed and evaluated with intelligibility in mind.

1.2 Research Questions

Based on the existing academic literature, which is laid out in greater detail in Chapter 2, this thesis aims to answer the following key research questions:

- R1: What type of driving scenarios primarily require explanations for AV passengers and what type of explanations are appropriate for these scenario types?
- R2: How can intelligible posthoc explanations of these types be generated automatically for common AV actions in the identified scenarios?
  - R2.1: How can we represent the core operations of an AV, and provide data structures, and algorithms for the generation of the relevant explanation types?
  - R2.2: How do we apply the algorithms from R2.1 on practical driving tasks?
- R3: How would passengers react to explainable but fallible autonomous driving systems?

Research question R1 aims to investigate driving scenarios obtained in the real world and identify the scenarios where explanations could be primarily useful. Driving scenarios are defined by a mix of road topology, road rules/traffic control signs, and vehicle actions with respect to other road participants. Explanations with and without causal attributions would be provided in the different scenarios.

Explanation intelligibility will be assessed based on the degree to which an explanation improves a person's understanding of an AV's operation after performing a set of tasks. In addition, accountability and trust factors would as well be assessed after participants have been conditioned to different explanations in different scenarios.

Research question R2 investigates how the relevant explanations identified from R1 can be generated programmatically for AV actions. This involves exploring ways to create enhanced AV architecture and data representation schemes that easily support effective explainability in AVs. Generating explanations requires the design of algorithms that can obtain the necessary explanation elements following these data representation schemes. These algorithms are expected to be able to provide intelligible explanations for specific prediction tasks in autonomous driving.

Research question R3 examines the effects of perception system errors on passengers, in an explainable AV that provides intelligible explanations of varying specificity. This is done through a lab study that employs a driving simulator with an immersive experience. The key to be investigated include perceived safety, the feeling of anxiety, and the feeling to takeover control.

While there may be some technology overlap between autonomous vehicles, uninhabited aerial vehicles (UAV) and autonomous underwater vehicles (AUV), we have only focused on autonomous vehicles (sometimes referred to as highly automated vehicles) in this thesis to enable us to cover enough depth.

There are a set of outcomes from this research. These outcomes are described in the next section.

1.3 Contributions

The thesis provides a number of contributions to research on explainable autonomous driving and human-machine interaction.

- **A comprehensive literature survey on explanations in autonomous driving**. It takes an interdisciplinary approach (sociotechnical) in that it establishes relationships between explainable AI theories, applications, and

human-machine interaction. Many existing works have only focused on vision-based explanation approaches in AVs. This survey was published in the IEEE Transactions on Intelligent Transportation Systems.

- **The first work to examine different dimensions of explanations: factual (why explanation), contrastive (why not explanation), counterfactual (what if explanation), and informative (what explanation) using the intelligibility, accountability, and trust goals in carefully selected highly diverse scenarios categorised into four: normative, near-miss, emergency, and collision.** While the previous body of work has explored this line of research, the dimensions of explanations and scenarios studied in this thesis are higher than those in the previous studies. It is also the first to assess explanations using the intelligibility, accountability, and trust objectives complementarily in autonomous driving. The outputs from this work were presented and published in the 2021 IEEE Intelligent Vehicles (IV) symposium and the 2021 IEEE Advanced Robotics and Its Social Impact (ARSO) conference.

- **Explainable AV representations and interpretable algorithms for generating posthoc approximate factual and counterfactual natural language explanations for collision risks and AV navigation actions.** While algorithms for generating natural language explanations have been proposed in previous works, the algorithms are mostly built using deep natural language processing architectures, which are in themselves not interpretable. Moreover, most of the existing textual explanation techniques have not offered counterfactual explanations in the autonomous driving context. The output from this work was presented and published in the 2021 IEEE Intelligent Vehicles (IV) symposium.

- **A novel dataset for advancing explainable AV research.** A multimodal (e.g., monocular dashboard camera and CAN Bus) and multilabel (e.g., agent

type, state, commentary, influences) driving dataset is provided, annotated with high-level semantics to facilitate explainable driving model designs.

- **A new tool for automatically providing natural language explanation labels for driving scenes in Carla simulator**. This tool has been demonstrated to different audiences, including audiences in the largest driving festival in the UK (i.e., the 2022 Goodwood Festival of Speed).

- **A new use case of varying levels of transparency through explanations in the presence of varying degrees of AV perception system errors.** While previous work has mainly focused on investigating the effect of explanations on end-users, we introduce a use case where explanation specificity and perception system errors in an AV are varied. A methodology that uses a psychometric scale, visual and haptic signals to assess the psychological effects of this set up on passengers was used. Findings are reported in this thesis.

- **An immersive driving simulation test bed with state-of-the-art virtual reality (VR) headset and explainer software for explainable autonomous driving research.** We have provided an immersive driving simulator with an explainer embedded for explainable autonomous driving research. The simulation environment is powered by Carla, a highly realistic autonomous driving simulator.

- **Regulatory suggestions for AV explainabiilty.** As this work is partly motivated by the existing regulatory framework on explanations (e.g., GDPR), we provided different regulatory suggestions building on the ICO and GDPR as a step towards achieving explainable AVs in the United Kingdom. This is necessary as there is currently an absence of comprehensive regulations or guidelines on explainability for autonomous vehicles.

1.4 Terminology

Throughout this thesis, we will use the terms:

- Intelligent Vehicle: any vehicle that can perform some or all driving operations with little or no input from a human driver.
- Autonomous Vehicle: any vehicle that can drive itself for a given time range with or without geographical restrictions.
- Intelligibility: the quality of being easily understood or comprehensible.
- Stakeholder: an individual or an agent whose roles involve direct or indirect interaction with an AV.
- Explanation: We note that there are different definitions of explanations in psychology (Dodwell, 1960), philosophy (Zalta et al., 1995) and AI (Ciatto et al., 2020; Omicini, 2020); we assume a more general meaning by referring to an explanation as 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 (Omeiza et al., 2022).
- Explainability: the ability of a system to support the provision of this form of explanation.
- Interpretable techniques: techniques that are transparent enough to support meaningful interpretations to their intended audience (mostly developers). We do not expect laypeople to be able to easily and quickly make sense of interpretable models (Poursabzi-Sangdeh et al., 2021). However, intelligible explanations (e.g., in natural language) for the outputs of the interpretable models will be beneficial to laypeople. This type of explanation is easier to obtain from interpretable models as shown in (Nahata et al., 2021; Stepin et al., 2021).

- Attributions or influence scores: methods that explain the decision of a black-box model by providing continuous values whose magnitude indicates the importance of each input feature for a given prediction (J. Zhou et al., 2021).
- Haptic signal: this is a signal obtained from tactile movements that involve muscles and joints.

1.5 Thesis Outline

The thesis is structured as follows: Chapter 2 provides a comprehensive survey of the existing body of work around explainable autonomous driving and human-machine interaction. First, we open with an overview of explainable AI in machine learning. We then provide motivations for explanations by highlighting and emphasising the importance of transparency, accountability, trust, and perceived safety in AVs; and examining existing regulations and standards. Second, we identify, describe and categorise the stakeholders involved in the development, use, and regulation of AVs. Third, we provide a thorough review of previous work on explanations for the different AV operations (i.e., perception, localisation, planning, control, and system management). Fourth, we discuss the research gaps that motivate this thesis.

Chapter 3 describes a study that elicits requirements for effective explainability in AVs. We describe the user studies we have conducted to explore explanation types in different autonomous driving scenarios.

Chapter 4 presents interpretable representations of AV operations and data to support effective explainability in autonomous driving. It also presents explanation generation algorithms that we have designed to explain collision risks and AV actions.

In Chapter 5, we describe experiments in which the designed algorithms were applied to explain collision risk models trained on the Lyft Level5 dataset, and action prediction models trained on our new explainable driving dataset (SAX dataset). We describe the data collection procedure carried out to obtain the SAX dataset needed for designing transparent explanation algorithms in autonomous driving. We describe the field trial and the data annotation process of the collected dataset.

We also present an interpretable benchmark model for explanation generation for AV actions.

In Chapter 6, we describe a laboratory study where we examined the effects of errors in AV perception models, exposed through different specificities of explanations. First, we state the hypotheses of interest, the study procedure and the assessment metrics. We present qualitative and quantitative results and discuss the implication of the results.

Chapter 7 concludes by discussing the contributions of the thesis as a whole, as well as highlighting associated limitations. We provide recommendations for addressing some of the identified challenges. Recommendations include regulatory suggestions.

As different methodologies were adopted, methodologies were discussed within each of Chapters 3 to 6.

1.6 Dissemination

The research described in this thesis has been and is being disseminated in the following venues.

1.6.1 Publications

- **Daniel Omeiza**, Raunak Bhattacharyya, Nick Hawes, Marina Jirotka, and Lars Kunze (2023). "Effects of Explanation Specificity on Passengers in Autonomous Driving". Under Review for the 2023 IEEE Intelligent Transport Systems Conference. Appeared in Chapter 6 of this thesis.

- **Daniel Omeiza**, Raunak Bhattacharyya, Marina Jirotka, and Lars Kunze (2023). "Effects of Explanation Specificity and Autonomous Vehicles' Perception System Errors on Passengers' Perceived Safety". For submission to Transportation Research Part F. Appeared in Chapter 6 of this thesis.

- **Daniel Omeiza,** Sule Anjomshoae, Helena Webb, Marina Jirotka, and Lars Kunze (2022). "From Spoken Thoughts to Automated Commentary Driving:

Predicting and Explaining Intelligent Vehicles' Actions." *In Proceedings of the IEEE Intelligent Vehicles Symposium.* doi:https://doi.org/10.1109/IV51971.2022.9827345. Appeared in Chapter 4 and 5 of this thesis.

- **Daniel Omeiza**, Helena Webb, Marina Jirotka, and Lars Kunze (2021). "Explanations in Autonomous Driving: A survey". *IEEE Transactions on Intelligent Transportation Systems.* doi:https://doi.org/10.1109/TITS.2021.312286. Appeared in Chapter 2 of this thesis.

- **Daniel Omeiza**, Konrad Kollnig, Helena Web, Marina Jirotka, and Lars Kunze (2021). "Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving". *In Proceedings of the IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO).* doi:https://doi.org/10.1109/ARSO51874.2021.9542835. Appeared in Chapter 3 of this thesis.

- **Daniel Omeiza**, Helena Webb, Marina Jirotka, Lars Kunze (2021). "Towards Accountability: Providing Intelligible Explanations in Autonomous Driving". *In Proceedings of the IEEE Intelligent Vehicles Symposium.* doi:https://doi.org/10.1109/IV48863.2021.9575917. Appeared in Chapter 3 and 4 of this thesis.

- Richa Nahata, **Daniel Omeiza**, Rhys Howard, and Lars Kunze (2021). "Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety". *In Proceedings of the IEEE 24th International Conference on Intelligent Transportation Systems (ITSC).* doi:https://doi.org/10.1109/ITSC48978.2021.9564966. Appeared in Chapter 4 of this thesis.

1.6.2 Other Publications Outside this thesis

- Marc Alexander Kühn, **Daniel Omeiza**, Lars Kunze (2023). "Textual Explanations for Automated Commentary Driving". 2023 IEEE Intelligent Vehicles Symposium.

- Pawit Kochakarn, Daniele De Martini, **Daniel Omeiza**, Lars Kunze (2023). "Explainable Action Prediction through Self-Supervision on Scene Graphs". IEEE International Conference on Robotics and Automation (ICRA).

- Lars Kunze, Omer Gunes, Dylan Hillier, Matthew Munks, Helena Webb, Pericle Salvini, **Daniel Omeiza**, and Marina Jirotka (2022). "Towards Explainable and Trustworthy Collaborative Robots through Embodied Question Answering". ICRA 2022 Workshop on the Collaborative Robots and the Work of the Future.

- Sule Anjomshoae, **Daniel Omeiza**, Lily Jiang (2021). "Context-based Image Explanations for Deep Neural Networks". *Image and Vision Computing.* Vol. 116. doi:https://doi.org/10.1016/j.imavis.2021.104310

- **Daniel Omeiza**, Sule Anjomshoae, Konrad Kollnig, Oana-Maria Camburu, Kary Främling, and Lars Kunze (2021). "Towards Explainable and Trustworthy Autonomous Physical Systems." *In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* doi:https://doi.org/10.1145/3411763.3441338

1.6.3 Demonstrations

- Omeiza et al. (2022). Automated Commentary Driving at Goodwood Festival of Speed. West Sussex, UK.

- Omeiza et al. (2022). Automated Commentary Driving at the Trust in Autonomous Systems All Hands Conference. London, UK.

2

Background & Literature Review

Contents

In this chapter, we provide a structured and comprehensive overview of the recent works on explainability in autonomous driving and provide a background for this thesis. Explanations have been studied in different domains. For example, explanations are considered to be useful in providing justifications in recommender systems (Bilgic & Mooney, 2005; Chang et al., 2016; Cleger et al., 2014; Gedikli et al., 2014; Guesmi et al., 2021; Hada & Shevade, 2021; Herlocker, 1999; Herlocker et al., 2000). In the general robotics domain, explanations have been offered as a means of facilitating human-robot collaboration and reconciling robot plans with human expectations (Chakraborti, Sreedharan, et al., 2017; Hoffmann & Magazzeni, 2019; Raman & Kress-Gazit, 2012; Raman et al., 2013; Sreedharan et al., 2019; Zhang et al., 2017). There are also previous works around the use of natural language models to provide explanations for domain-specific tasks, e.g., in image captioning (Hendricks et al., 2016; Hendricks et al., 2018). In this literature review, we focused on works related to explainable autonomous driving while drawing insights from other applications. Previous literature survey papers have focused on approaches aimed at 'opening' black-box machine learning mechanisms (data-driven XAI) applied in deep learning (Adadi & Berrada, 2018a; Guidotti et al., 2018; Samek et al., 2017a). In contrast, Anjomshoae et al. (2019) provided a systematic literature review generally on explainable agencies (i.e., explaining the behaviour of goal-driven agents and robots) which entailed the use of descriptive statistics to show the amount of research done around explainable agencies with no particular focus on autonomous driving. A related work surveyed the literature around explanations in vision-based autonomous driving systems (Zablocki et al., 2021). This chapter aims to fill the gap in the academic literature by providing a comprehensive survey on explanations for the behaviour of AVs at different aspects of operations (i.e., perception, localisation, planning, vehicle control, and system management) with the requirements of different stakeholders in mind. The survey also identifies research gaps which are addressed in the subsequent chapters of this thesis.

The rest of this chapter is organised into 8 sections: Section 2.1 provides a general overview of explainable AI in machine learning. Section 2.2 presents the

general need for explanations in autonomous vehicles. Section 2.3 presents and discusses the regulations and standards related to explanations in AVs. The different stakeholders who interact with AVs are identified and categorised in Section 2.3.3. The categorisation system defined in Section 2.3.3 is used in the rest of the chapter. Section 2.4 broadly categorises explanations into many dimensions and provides several literature references for the different categories. Section 2.5 describes the core operations of an AV and reviews existing work on explanations in relation to the different core AV operations. These operations include perception, localisation, planning, and control. Section 2.6 examines AV system management. System management involves event data recorders and human-machine interaction, which are crucial for explanations. Section 2.7 summarises the research gaps.

2.1 Overview of Explainable AI in Machine Learning

Explainable AI (XAI) 'is a research field that aims to make AI systems results more understandable to humans' (Adadi & Berrada, 2018b). The primary focus of XAI within the realm of machine learning pertains to creating machine learning models that are capable of presenting comprehensible justifications for their predictions to humans. It also involves the creation of methodologies to enable the models to do so. The explanations generated from XAI methods can come in different forms, e.g., feature importance, feature attributions (such as saliency, attention, natural language) and can be useful in debugging and improving the performance of machine learning systems (Samek et al., 2017b). As described in Adadi and Berrada (2018b), explainability in machine learning is conceptualised to be either intrinsic or post-hoc. It is intrinsic when the system themselves are transparent, and it is post-hoc if the system concerned is opaque or blackbox and requires a different system to provide an explanation. Explainers (i.e., explanation generation techniques) can be model-specific or model agnostic. Further, some explainers are only able to provide explanations for individual predictions, while some can provide explanations about an entire model. These are referred to local and global explanations, respectively.

2.1.1 Intrinsic Explainability

This concept of explainability is hinged on the underlying model's complexity. Machine learning models with simple structure, architecture or a few parameters are considered intrinsically interpretable as they do not require additional models to explain their behaviour and decisions (Ai & Narayanan. R, 2021; Das & Rad, 2020). Examples of intrinsically interpretable models are decision trees and linear models.

Decision Trees

A decision tree model is a non-parametric predictive model that draws conclusions about observations using supervised learning (Shalev-Shwartz & Ben-David, 2014). Decision trees can be used for regression tasks (target variable takes continuous values) and classification tasks (target variable takes a discrete set of values). A decision tree model follows a binary tree structure with a root node (node with no incoming edge), internal nodes (nodes with both incoming edge and outgoing edges), and leaf nodes (nodes without outgoing edges). An edge is the connection between two nodes.

The tree is built in the learning phase, where the source set is divided into subsets based on a set of features in the dataset. These subsets become the successor children, with the original set serving as the root node. A set of rules guides this splitting process, and it is repeated recursively on each derived subset, a process called recursive partitioning. The recursion stops when a subset at a node is homogeneous or when further splitting does not improve the accuracy of predictions. Splitting rules can be formed using entropy, information gain, or gini impurity measures, among others (Suthaharan, 2016). A process called pruning (Marsland, 1986) can be done on the tree to prevent overfitting; the depth of the tree can also be constrained to reduce tree complexity.

Intelligible explanations from decision trees: There are methods to visualise a tree model in order to trace decision paths (Parr & Grover, 2020). Some methods translate the split conditions in the decision path to high-level semantics that

lay users can understand (Stepin et al., 2021). Further, local feature importance scoring technique was proposed in Palczewska et al. (2013) for estimating the local importance scores of features for a prediction. While these are useful techniques, there still exists the challenge of constraining the length of explanations for a prediction when the tree is deep. Moreover, generating counterfactual explanations (i.e., *What Ifs*) under defined constraints for tree models is yet to be explored.

Multiple decision trees can be created during training, a process called ensemble learning. The resulting model is called a Random Forest (Breiman, 2001) which is not intrinsically interpretable, but less complex and more transparent than deep neural network models. This is due to the complexity of its decision making process. For classification tasks, the class with the highest number of votes from the trees is selected as the output. Meanwhile, the mean or median prediction of all trees is returned as the output for regression tasks. To the best of our knowledge, our work is the first to develop a robust tree-based explainer for autonomous driving tasks.

2.1.2 Post-Hoc Explainability

This refers to the process of explaining the decision-making process of a machine learning model after a decision or prediction has been made. Post-hoc explainability aims to provide insights into how the model or AI system arrived at its conclusions. A post-hoc explainability method can either be model specific or model-agnostic.

Model Agnostic Explainers

Model agnostic explainers are used to explain the predictions or decisions made by any machine learning model, irrespective of the structure of the model. These methods usually do not require access to the internal workings of the model but rather use a proxy model to approximate the behaviour of the original model by inspecting the models' output around inputs within a certain neighbourhood. Model agnostic explainer could be local or global.

Local Explainers: Local explainers provide explanations for individual predictions made by machine learning models. These explanations focus on understanding how a specific prediction was arrived at rather than providing a general explanation for the entire model. Examples are SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), among others.

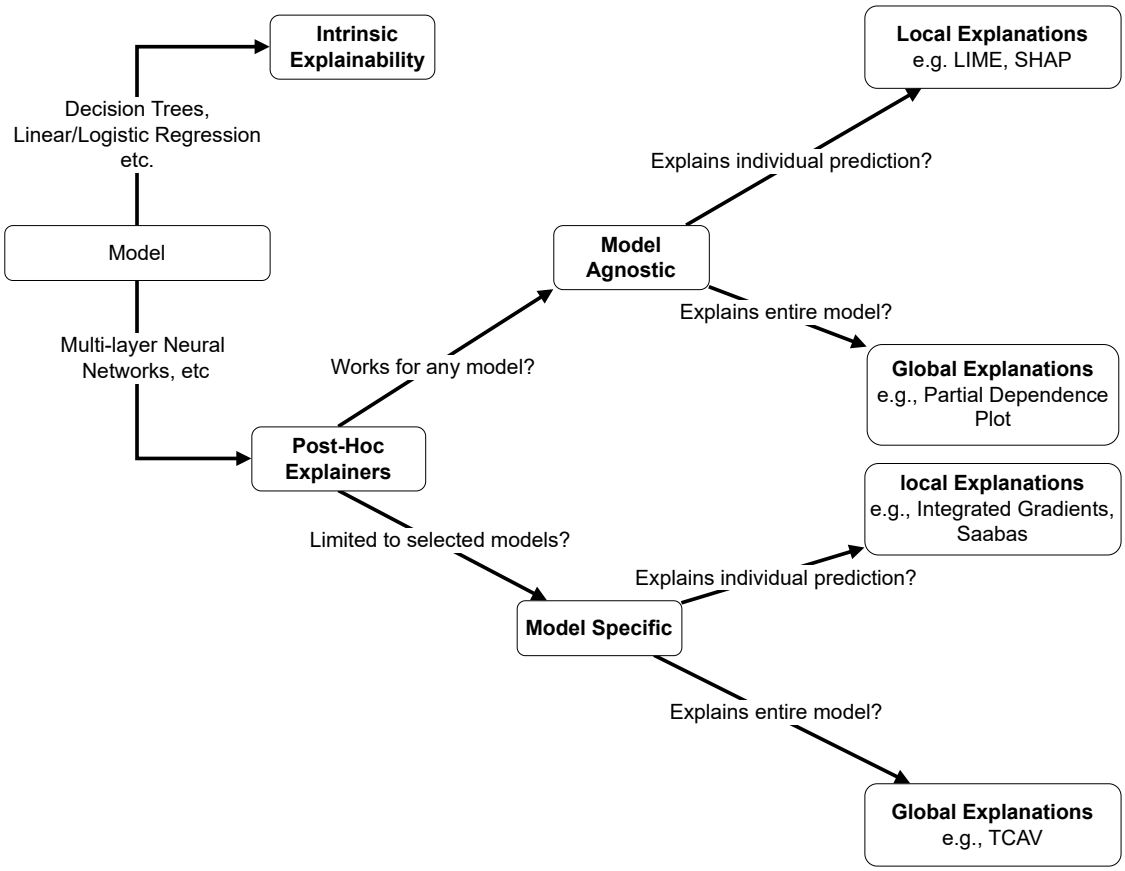


Figure 2.1: Explainable AI techniques in machine learning

Global Explainers: Global explainers provide explanations for the overall behaviour of a machine learning model rather than focusing on individual predictions. These methods aim to provide a holistic understanding of how the model works, including which features or variables are most important in making predictions and the relationship between these variables. Examples of this type of explainer are Tree SHAP (Lundberg et al., 2020), and the partial dependence plots (Greenwell, 2017).

Global explainers are particularly useful in applications where understanding the overall behaviour of a machine learning model is important, such as in regulatory compliance, auditing, or risk management.

2.1.3 Model Specific Explainers

Model specific explainers can only provide explanations for specific types of machine learning models. Model specific explainers could either be local or global as well. Examples of local model specific explainers are the gradient based attribution methods e.g., Grad-CAM (Selvaraju et al., 2017), Integrated Gradient (Sundararajan et al., 2017), Saabas (Saabas, 2014). An example of a global and model specific explainer is the Testing with Concept Activation Vectors (TCAV) explainer (B. Kim et al., 2018).

2.2 Need for Explanations in Autonomous Driving

The need for explanations in autonomous vehicles stems from the increasing concerns for transparency, accountability, safety, and trust in autonomous vehicles. It is believed that explanations are one way of achieving these goals. In this section, we discuss the need for explanation in light of transparency, accountability, safety, and trust.

2.2.1 Transparency and Accountability

One generally agreed-upon notion of accountability is associated with the process of being called 'to account' to some authority for one's actions (Jones, 1992). Accountability, in broad terms, often encompasses closely related concepts, such as responsibility and liability (Martinho et al., 2021). Mulgan (Mulgan, 2000) elucidated that accountability entails responsibility but, unlike responsibility, it requires explanations about actions and cannot be shared. Meanwhile, liability is a legal or financial responsibility (Collingwood, 2017). In the human and machine context, Doshi-Velez et al. (2017) conceptualise accountability as the ability to

determine whether the decision of a system was made in compliance with procedural and substantive standards, and importantly, to hold one responsible when there is a failure to meet the standards. In autonomous driving, accountability becomes a challenging issue mainly because of the various operations involved (e.g., perception, planning, controls, and system management among others) that demand inputs from multiple stakeholders; this can result in responsibility gaps.

As identified by (Mulgan, 2000), achieving accountability requires social interaction and exchange. At one end, the requester of an account seeks answers and rectification while at the other end, the respondent or explainer responds and accepts responsibility if necessary. In the context of this review, the AV is being called by a stakeholder to provide an account; one expects the AV to provide an account in the form of an explanation that is intelligible to the requester to facilitate the assignment of responsibilities. There have been debates on how responsibility should be allocated for certain AV accidents. Companies have stated the need for clear rules to be set in advance. For example, Honda has reported that it is necessary to put legal frameworks in place in order to clarify where the responsibility lies in case of the occurrence of an accident after the realisation of fully automated driving ("Honda sustainability report (Tech. Rep.)" n.d.). Technical solutions are also being put forward. One such example is the proposal for the use of a 'blackbox', similar to a flight recorder in an aircraft, to facilitate investigations ("Sustainable value report (Tech. Rep.)" n.d.). Shashua and Shalev-Shwartz (Shashua & Shalev-Shwartz, 2017) also advocated for the use of mathematical models to clarify faults in order to facilitate a conclusive determination of responsibility.

The social aspect of accountability described by (Mulgan, 2000), will demand that the aforementioned recommended approaches are able to plug into explanation mechanisms where causes and effects of actions can be communicated to the relevant stakeholders in intelligible ways. In addition to accountability for accident cases, which has gained much attention in industry reports, actions resulting in undesired, discriminatory, and inequitable outcomes also need to be accounted for. This means that stakeholders such as passengers or auxiliary drivers who may not have direct

involvement in the management of the AVs should be able to instantaneously request accounts as intelligible explanations for such undesired actions when they occur.

2.2.2 Perceived Safety in AVs

Safety is often referred to as a situation with a lower risk compared to an acceptable risk or a situation 'without any danger impending' (Ebi, 2009). Safety could either be objective, that is, based on an objective evaluation of the safety factors, or could be subjective, that is, safety based on feeling or perception (Z. Li et al., 2013). Hence, the term perceived safety or perception of safety as used in this thesis.

Safety is considered important for trust building process in automation. This is even more critical in autonomous vehicles. In a public survey (Jardim et al., 2013), the importance of safety, costs, and laws on respondents' perception of AVs were gathered. Safety was ranked (by the respondents) as the most important aspect to consider before adopting AVs.

While autonomous vehicles promise many benefits, factors such as the sudden deviation from the norm of humans directly taking charge of navigation operations make the public more hesitant towards this nascent technology. Moreover, accident reports associated with highly automated vehicles (Lavrinc, 2016; McFarland, 2016; NTBS, 2018; Stanton et al., 2019; "Tesla deaths", n.d.; Tilley, 2016) might hamper the feeling of safety and trust (Yurtsever et al., 2020b). J. Wang et al. (2020) studied 128 accidents that occurred between 2014 and 2018. About 63% of the accidents were caused in autonomous mode. However, only 6% of the total accident were directly related to AVs. Even when AVs become highly safe, the perception of safety of the public might still remain unchanged for a period of time.

Improved human-machine interfaces are believed to bridge this gap between humans and vehicle technology. Users may feel safer in vehicles that act more human-like. When autonomous vehicles act in ways that are more machine-like, such as acting more assertively in congested areas due to their logic, the user is more likely to feel unsafe or uneasy (Oliveira et al., 2019). Section 2.6 discusses the recent research developments around HMIs and explanations.

2.2.3 Trust

M. Faas et al. (2021) argued that research investigating trust in automation has been around for decades, i.e., since the introduction of interpersonal trust theories into the human-machine interaction domain by (J. Lee & Moray, 1992; M. Faas et al., 2021; Muir, 1987, 1994). While various definitions of trust in automation have been proposed, the most commonly adopted definition is that put forward in (J. D. Lee & See, 2004). The authors consider trust as a social psychological concept that is important for understanding automation partnerships. The definition stresses that trust is the attitude that an agent or automation will help an individual to achieve their goals in a situation characterised by uncertainty and vulnerability. Trust in automation, as made evident in (Biros et al., 2004; Hergeth et al., 2016; Muir & Moray, 1996), has significantly influenced the acceptance of and reliance on automated systems. As opposed to a binary categorisation, trust can be more finely calibrated so that an individual's trust levels in an automated system adequately reflect the actual capabilities and functional scope of an automated system. This trust calibration is considered to be an important requirement for safe and efficient human-machine interaction (J. Kraus et al., 2020; Muir, 1987). While calibration is useful, miscalibrated trust is disastrous as it can lead to distrust or overtrust (i.e., excessive trust). This will either make the user underuse the system or use the system beyond the scope of its functionalities (M. Faas et al., 2021). The process of using available information to assess and learn about the trustworthiness of a system to adapt trust levels is referred to as trust calibration (Khastgir et al., 2018; J. Kraus et al., 2020; J. M. Kraus, 2020).

Information about the functioning modes of an AV at the user's disposal can help the user create a better understanding of the AV's behaviour, eventually adding to the user's knowledge base (Hoffman et al., 2018), and helpful for constructing calibrated trust. This information could be presented as explanations of the operational modes and behaviour of a complex system, such as an AV, especially when it acts outside the expectations of the user. It is noted that trust can break down when there are frequent failures without adequate explanations, and regaining

trust once lost can be challenging (Dzindolet et al., 2003; Madhavan & Wiegmann, 2007). For example, previous reports on AV accidents may have a negative impact on calibrated trust in AVs. According to Hussain and Zeadally (2018), a serious challenge evident in intelligent transport systems is the lack of trust from the consumer's perspective. The public fears that the claims on accidents reduction through the introduction of AVs may be misleading as they consider human drivers to be better than AVs in handling (Hussain & Zeadally, 2018) unforeseen and uncharacteristic traffic situations. Abraham et al. (2016) also reported that the consumers' perception of trust is still not as high as expected in spite of the great potential promised by AVs, claiming that the public is still hesitant about the technology, and still feel uncomfortable using it. Trust is, therefore, imperative for achieving widespread deployment and use of AVs.

Researchers, e.g., (Hoffman & Klein, 2017), suggest that the provision of meaningful explanations from AVs to stakeholders (e.g., passengers, pedestrians and other road participants) is one way to build the necessary trust in AV technology. Other empirical studies (Ha et al., 2020; Koo et al., 2015; Omeiza, Kollnig, et al., 2021) have shown that the provision of explanations in AVs can influence trust. While it has been argued in (Hergeth et al., 2016; Payre et al., 2016; Rajaonah et al., 2006) that trust is a substantial subjective predicting factor for the adoption of automated driving systems, several studies have shown the importance of viewing trust formation and calibration in AVs as a temporal process influenced by prior information or background knowledge (Beggiato & Krems, 2013; J. M. Kraus et al., 2019). Explanation provision in autonomous driving over time is therefore crucial. In the following section, we will discuss explanations from the regulatory perspective.

2.3 Regulations, Standards, and Stakeholders

2.3.1 Guidelines and Regulations

We restrict the scope of this section to only relevant regulations and guidelines in Europe. There are increasing concerns about the collection and use of personal data in algorithms that make critical decisions about people in domains like healthcare,

finance, insurance, and criminal justice. The European Union GDPR implemented in 2018 aims to provide more control rights to individuals over their personal data ("A right to explanation", n.d.). The GDPR also sets guidelines related to the explanation of decisions made based on users' data. The GDPR guideline mandates that controllers (entities handling people's personal data) provide meaningful information about the logic involved in the decisions made based on people's data and what the likely consequences are for individuals. It also demands the appropriate use of mathematical or statistical procedures on such data. This is commonly referred to as the 'right to explanation'. In addition, the GDPR Article 12, which stresses transparency, demands that the provision of information/explanation to data subjects must be done in an intelligible way, i.e., in a clear and easily understandable form. These clauses highlight the user's right to question the decision of a system and the demand for explanations, especially when decisions are made based on their data.

The UK sets an ethics, transparency and accountability framework for automated decision-making (GOV.UK, n.d.). In this framework, it states that:

> When automated or algorithmic systems assist a decision made by an accountable officer, you should be able to explain how the system reached that decision or suggested decision in plain English. The explanation needs to be appropriate for your audience, expert or non-expert and should be scrutinised and iterated by a multidisciplinary and diverse team (including end-users) to avoid bias and group speak.

The ICO in the UK provided organisations with 'practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them.' (Information Commissioner's Office, n.d.). The ICO—in their guidelines—identifies two subcategories of explanation (*process-based* and *outcome-based* explanations) and urges organisations to consider these subcategories when providing information to subjects. Process-based explanations provide information on the governance of the organisation's AI system across its design and deployment; while outcome-based explanations relay information concerning a specific outcome resulting from a decision made by said system.

Processed-based explanations, as defined by the ICO, transcend global explanations as they include the system governance process, while outcome-based explanations can be likened to local explanations.

An autonomous vehicle can potentially be used to collect sensitive information from users either legally or illegally. By tracking an AV, a passenger's location is known, a passenger's frequent routes can be determined, as well as the time of the day they typically travel. Hence, autonomous vehicles should not be exempt from these general explainability clauses, especially when they operate in the regions where this regulation holds.

Closely related to autonomous vehicles is the 2020 ethics of connected autonomous vehicles (CAV) report provided by an expert group convened by the European Commission (Bonnefon et al., 2020). Explainability recommendations were made for manufacturer and developers:

> Manufacturers and deployers should develop and implement user-centred methods and interfaces for the explainability of relevant CAV applications of algorithm and/or machine learning based operational requirements and decision-making. They should ensure that the methods and vocabulary used to explain the functioning of CAV technology are transparent and cognitively accessible, the capabilities and purposes of CAV systems are openly communicated, and the outcomes traceable. This should ensure that individuals can obtain factual, intelligible explanations of the decision-making processes and justifications made by these systems, particularly in the event of individually or group-related adverse or unwanted consequences.

Recommendations were also made for researchers:

> Researchers should aim to develop explainability-enhancing technologies in relation to data collection and algorithms used for CAV decision-making. They should formulate methods for designing CAV systems which guarantee that datasets and algorithms are thoroughly documented, meaningfully transparent and explicable in a way that is adapted to the expertise of the parties concerned (e.g., individual users, policymakers, etc.) More broadly, further empirical, technical, normative/philosophical and legal research is needed to explore methods and safeguards of explainable AI that help to mitigate against biases and discrimination risks.

There are other regulations and guidelines being set for autonomous vehicles in Europe, for example, the Scottish Law Commission's regulation for autonomous vehicles (Scottish Law Commission, n.d.), and the preliminary consultation paper on autonomous vehicles (Law Commission, n.d.) by the UK Law Commission.

While the regulatory recommendations and guidelines are quite abstract with respect to autonomous vehicles requirements and implementations, key takeaway for this research is that (i) AVs should be explainable (ii) they should be able to provide factual (causal) explanations that are intelligible to lay users (iii) research efforts are needed to clearly define effective requirements for AV explainability, and formulate methods that facilitate explainability and transparency in general. These are key motivations for this research.

2.3.2 AV Standards

Intelligent Transport Systems (ITS) apply advanced electronics, information and communications technologies to roads and automobiles. This is done to collect, store, and provide traffic information in real-time for convenient and safe transport; improved reliability, efficiency, and quality; and the reduction in energy consumption ("Intelligent transport systems", 2018). The International Standard Organisation Technical Committee 204 (ISO TC204), the IEEE, and related standard organisations have set standards for AVs and ITS in general. The IEEE Initiatives, in particular, has a vision for prioritising human well-being with autonomous and intelligent systems, and the assessment of gaps in standardisation for safe autonomous driving. The very recent IEEE P7001 standard ("IEEE Standard for Transparency of Autonomous Systems", 2022) was motivated by the possibility of understanding why and how an autonomous system behaved. It aimed to set out objective measurements for transparency in autonomous systems in general. These standards directly or indirectly demonstrate the necessity of explainability in AVs. In Table 2.1, we identified standards that are related to safety and information/explanation provision in AVs. We categorised these standards into two sets: Human safety-related standards and information or data exchange related

standards. Further details on AV related standards are available in the ISO report on intelligent transport systems in ("World Report for Intelligent Transport Systems (ITS) Standards - A Joint APEC-International Organization for Standardization (ISO) Study of Progress to Develop and Deploy ITS Standards (ISO TR 28682)", 2017) and the Apex.AI document on automated mobility ("An overview of taxonomy, legislation, regulations, and standards for automated mobility", 2020).

I describe the different stakeholders concerned in AV explanations in the next section.

2.3.3 Stakeholders

Explanation provision in autonomous driving has many personas due to the different purposes of explanations. The level of detail (in terms of information) anticipated by the explanation recipients, the explanation type and the mode of communication vary with respect to the type of recipient and purpose for the explanation. This highlights the importance of explanation personalisation with respect to stakeholders. Personalisation is seen to be crucial for the generation of intelligible or understandable explanations (Kouki et al., 2019; Meske et al., 2020; Shin, 2021). While lay users who lack technical domain expertise may be satisfied with a user-friendly explanation that requires less background knowledge to interpret, developers and engineers would prefer a finely detailed explanation with technical terms that would support a deeper conception of the internal functioning of a model (Y. Zhou & Danks, 2020). In this light, the consideration of the persona of the explainee is necessary (Langley, 2019). Going forward, we refer to anyone who has to engage with an explanation as a stakeholder. Having identified the typical personas in the literature, we divided stakeholders into three broad categories: Class A (all types of end-users and society), Class B (all technical groups, e.g. developers), and Class C (all forms of regulatory bodies including insurers). See the further description:

1. **Class A: End-Users**

Table 2.1: Selected standards for autonomous vehicles. These standards underline the importance of safe, transparent, and explainable AVs.

Aim	Standard & Description	Stakeholder
Human Safety	**ISO 19237:2017** Pedestrian detection and collision mitigation systems	**Class B and C** AV Developers, Regulators, System Auditors, Accident Investigators, Insurer
	ISO 22078:2020 Bicyclist detection and collision mitigation systems	
	ISO 26262:2011: Road vehicles – Functional safety. An international standard for functional safety of electrical and/or electronic (E/E) systems in production automobiles (2011). It addresses possible hazards caused by the malfunctioning behaviour of E/E safety-related systems, including the interaction of these systems.	
	ISO 21448:2019: Safety Of The Intended Functionality (SOTIF). Provides guidance on design, verification and validation measures. Guidelines on data collection (e.g. time of day, vehicle speed, weather conditions) (2019). (complementary to ISO 26262).	
	UL 4600: Standard for Safety for Evaluation of Autonomous Products. a safety case approach to ensuring autonomous product safety in general, and self-driving cars in particular.	
	SaFAD: Safety First for Automated Driving. White paper by eleven companies from the automotive industry and automated driving sector about frameworks for the development, testing and validation of safe automated passenger vehicles (SAE Level 3/4).	
	RSS (Intel) / SFF (NVIDIA): Formal Models & Methods to evaluate safety of AV on top of ISO 26262 and ISO 21448 (proposed by companies).	
	IEEE Initiatives: "Reliable, Safe, Secure, and Time-Deterministic Intelligent Systems (2019)"; "A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" (2019); "Assessment of standardization gaps for safe autonomous driving (2019)".	
	The Autonomous: Global safety reference, created by the community leading automotive industry players, which facilitates the adoption of autonomous mobility on a grand scale (2019).	
Information/ Data Exchange	**ISO/TR 21707:2008: Integrated transport information, management, and control—Data quality in intelligent transport systems (ITS)**. "specifies a set of standard terminology for defining the quality of data being exchanged between data suppliers and data consumers in the ITS domain" (2018).	**Class A and C** Passengers, Auxiliary Drivers, Pedestrians, Regulators, System Auditors, Accident Investigators Insurers
	ISO 13111-1:2017: The use of personal ITS station to support ITS service provision for travellers. "Defines the general information and use cases of the applications based on the personal ITS station to provide and maintain ITS services to travellers including drivers, passengers, and pedestrians" (2017).	
	ISO 15075:2003: In-vehicle navigation systems—Communications message set requirements. "Specifies message content and format utilized by in-vehicle navigation systems" (2003).	
	ISO/TR 20545:2017: Vehicle/roadway warning and control systems. "Provides the results of consideration on potential areas and items of standardization for automated driving systems" (2017).	
	ISO 17361:2017: Lane departure warning.	
	ISO/DIS 23150: Data communication between sensors and data fusion unit for automated driving functions.	

- Passenger: this is the in-vehicle agent who may interact with the explanation agency in the AV but is not responsible for any driving operation.
- Auxiliary Driver: This is a special in-vehicle passenger who may also interact with the explanation agency in the AV and can also participate in the driving operations. This kind of participant may mainly exist in SAE level 3 and 4 vehicles.
- Pedestrian: this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI).
- Pedestrian with Reduced Mobility (PRM): this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) but have reduced mobility capacity (e.g., pedestrian in a wheelchair).
- Other Road Participants: these are other agents outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) (e.g., cyclists, other vehicles).

2. **Class B: Developers and Technicians**

- AV Developer: the agent who develops the automation software and tools for AVs.
- Automobile Technicians: the agent who repairs and maintains AVs.

3. **Class C: Regulators and Insurers**

- System Auditor: the agent who inspects AV design processes and operations in order to ascertain compliance with regulations and guidelines.
- Regulator: the agent who sets guidelines and regulations for the design, use, and maintenance of AVs.

- Accident Investigator: the agent who investigates the cause of an accident in which an AV was involved.
- Insurer: the agent who insures the AV against vandalism, damage, theft, and accidents.

In the next section, we provide a categorisation of explanations based on methodologies and situate the different stakeholders in the categorisation.

2.4 Explanation Categorisations from the Research Literature: A Broader View

Explanations serve different functions in different contexts (Y. Zhou & Danks, 2020). Therefore, the methods of generation and evaluation are context and purpose-dependent (Binns et al., 2018). D. Wang et al. (2019) identified three approaches that have been adopted in the academic literature in either developing or evaluating explanations.

First, the authors highlighted the existence of **unvalidated guidelines** for the design and evaluations of explanations. They claim that these kinds of guidelines are based on authors' experiences with no further substantial justification. Hence, explanation generation algorithms that generate explanations as short rules (Lakkaraju et al., 2016), or those that apply attributions or influence scores (Selvaraju et al., 2017)—such as partial dependence plots (Greenwell, 2017)—without sufficient justification for the explanation choices made are assumed to be based on unvalidated guidelines. Thus, the explanations generated by these algorithms may not be appropriate for class A stakeholders due to the low intelligibility quality of the explanations (Y. Zhou & Danks, 2020).

Second, researchers suggested (Zhu et al., 2018) that understanding users' requirements might be helpful in explainable AI research. It is on this premise that some research on explanation design approaches has been thought to be **empirically derived**. This type of research elicits explanation requirements from user surveys in order to determine the right explanation for a use-case with

explanation interfaces (D. Wang et al., 2019). For instance, explanation frameworks have been proposed for recommender systems (Herlocker et al., 2000), case-based reasoning (Roth-Berghofer, 2004), intelligent decision aids (Silveira et al., 2001), and intelligible context-aware systems (Lim & Dey, 2009) upon the elicitation of users' requirements through surveys and user studies. Through user studies, Lim and Dey (2009) examined explanations based on intelligibility types. The intelligibility types used were: 'why' (factual), 'why not' (contrastive), 'what if' and 'how to' (counterfactual) explanations which are considered relevant for filtering causes for an effect. We interchangeably refer to these intelligibility types as causal filters or investigatory queries in this thesis.

Third, some explanation design methods are derived from **psychological constructs from formal theories** in the academic literature (D. Wang et al., 2019). Some of these methods, e.g., in (Hoffman & Klein, 2017), draw on philosophy, cognitive psychology, social science, and AI theories to inform explanation design for explanation frameworks. For example, Akula et al. (2019) employed the Theory of Mind (ToM) in the development of an explanation framework (X-ToM). The authors in (Akula et al., 2019) claimed that in their explanation framework, the mental representations in ToM were incorporated to learn an optimal explanation policy that took into account human perception and beliefs. Simply put, a policy, as used in this context is an agent's strategy for achieving a goal (Sutton & Barto, 2018). Theory of mind involves explaining people's behaviour on the basis of their minds: their knowledge, their beliefs, and their desires (Frith & Frith, 2005). It is noted that there are criticisms of the theory of mind and mental models. However, this is out of the scope of this thesis.

I use the three discussed methodologies as one of the categorisation dimensions. Explanations methods that are mainly based on the researcher's experience without further user studies to justify claims are categorised under unvalidated guidelines (UG). Those that adopted a user study to elicit users' experience are categorised as empirically derived (ED), and those that built on psychology theories are categorised under psychological constructs from formal theories (PC). Other dimensions for

Table 2.2: Causal filters and example investigatory queries.

Causal Filter	Class	Example Query
Why Not (Contrastive)	Causal	why did you not do Y?
Why (Factual)	Causal	why did you do X?
What If (Counterfactual)	Causal	what would you do if Z?
What	Non-Causal	what are you doing?

categorisation include causal filter, explanation style, interactivity, dependence, system type, scope, stakeholders, and operation.

The description of the various dimensions of explanations is detailed below.

Causal Filters: explanations resulting from causal filters use selected *causes* relevant to interpreting an observation, with respect to existing knowledge (D. Wang et al., 2019). The explanations provided in this category are assumed to be usually generated by causal filters or investigatory queries like *why, why not, how to,* and *what if* (Lim et al., 2009). These causal filters are assumed to produce explanations that could be factual (e.g. 'why' explanation), contrastive ('why not' explanation), or counterfactual ('how to' and 'what if' explanation). See Table 2.2.

Explanation Style: explanations are categorised based on the type of information or elements referenced in the explanation and the forms they are presented in (Binns et al., 2018).

- Input Influence: a list of input variables is presented along with quantitative measures of their influence (either positive or negative) on a decision.
- Sensitivity: shows what magnitude of change is required in an input variable in order to change the output class. Note that this is different from the sensitivity used in machine learning evaluation.
- Case-based: picks out a relevant case from the model's training data that is most similar to the decision made, which is then used to explain.
- Demographic: explanation provides aggregate statistics of previous outcomes for people with the same demographics.

Model Dependence: in this context, it refers to the possibility of having an explanation method that can be used to explain any type of autonomous driving model (e.g., perception models and motion planning models). If the possibility exists, the explanation method is considered to be *model agnostic*. This is similar to the model specific and model agnostic explainability introduced in Section 2.1, but now contextualised to autonomous driving. Otherwise, it is regarded as *model specific*. Two popular model-agnostic explanation techniques are SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016). Although LIME and SHAP explanation techniques can be useful in autonomous driving, to the best of our knowledge, only SHAP has been used in the context of autonomous driving (Nahata et al., 2021).

Interactivity: this refers to the possibility of a stakeholder raising follow-up questions as a way of demanding further explanations. The conversational style of explaining (Miller, 2019) allows for this.

System Type: this refers to the nature of the system that the explanation technique is primarily designed for. It could be an explanation technique for *data-driven* systems (e.g., explaining the output of a machine learning model) or a *goal-driven* system (e.g., explaining the behaviour of an autonomous agent based on plans and goals) (Anjomshoae et al., 2019). In more detail, an explanation method that explains a deep learning model trained on driving scene images or video is data-driven while one that explains plans (or changes in plans) and/or actions with reference to a goal is referred to as goal-driven in this context.

Scope: in this context refers to the coverage of the explanation in terms of the system's parts. We adapt terminologies from the explainable AI (XAI) in machine learning domain. As in Section 2.1, a global explanation explains a model's overall behaviour or decision-making process, while a local explanation explains a single prediction (Lundberg et al., 2020; van der Linden et al., 2019). The term global explanation in this review is used to refer to an explanation that explains the entire behaviour of an AV. In contrast, a local explanation refers to an explanation that

only explains a subset of the AV's behaviour. Nahata et al. (2021) proposed a tree explanation technique that can provide both factual (*why*) and counterfactual (*what if*) explanations for an AV collision risk model. Users can specify simple constraints for generating counterfactual explanations (e.g., setting the desired counterfactual output to be explained).

A representative subset of previous works where an explanation technique was primarily discussed or implemented in the context of autonomous driving is shown in Table 2.3. Note that Chakraborti, Sreedharan, et al. (2017) and Raman et al. (2013) are mainly on robot plan explanations but apply to autonomous vehicles. While attention maps are commonly regarded as explanations in the machine learning literature, Jain and Wallace (2019) argued against this notion by claiming, based on the outputs of experiments, that attention maps are not explanations. Consequently, Wiegreffe and Pinter (2019) disproved this claim and argued that such a claim depends on one's definition of explanation and that prior work against the effectiveness of attention maps for explanations does not disprove the usefulness of attention mechanisms for explainability. We agree that attention maps and heatmaps are not effective in some cases but are however useful. Therefore, we include relevant works on attention maps and heatmaps in this survey. See Table 2.3 for an overview.

The overview provided in Table 2.3 indicates that some types of explanations (e.g., sensitivity, demographics, contrastive, counterfactual, model-agnostic, and global explanations) are rare in the autonomous driving literature. This may be due to the nascent nature of the explainable autonomous driving domain.

2.5 Explainable Autonomous Driving Operations

This section provides a high-level description of the different operations of an AV and a review of previous work on explanations related to each of the operations. The operations include perception, localisation, planning, control and navigation, and system management (which includes event data recorder and human-machine interaction) (Jo et al., 2014).

Table 2.3: Summary of explanations categories. The table includes a subset of the reviewed papers where each or a subset of the explanation categories was mentioned in the context of autonomous embodied agents.
Stakeholders: Class A—Passenger (PA), Pedestrian (PE), Pedestrian with Reduced Mobility (PRM), Other Road Participants (ORP), Auxiliary Driver (AD). Class B—Developer (DV), Auto-Mechanic (AM). Class C—System Auditor (SA), Regulator (RG), Insurer (IN), Accident Investigator (AI).
Methods: Unvalidated Guidelines (UG), Empirically Derived (ED), Psychological Constructs from Formal Theories (PC).
Operations: Perception (P), Localisation (L), Planning (PL), Control (C), System Management (M)

Ref.	**Causal Filter**			**Explanation Style**				**Interactivity**		**Dependence**		**System**		**Scope**		*Method*	*Stakeholders (Class)*	*AV Operations*
	Factual	*Contrastive*	*Counterfactual*	*Input Influence*	*Sensitivity*	*Case-based*	*Demographic*	*Conversational*	*Non-conversational*	*Model Agnostic*	*Model Specific*	*Goal-Driven*	*Data-Driven*	*Local*	*Global*			
J. Kim et al. (2018)	✓					✓			✓		✓		✓	✓		UG,ED	B, C	P, C
Chakraborti, Sreedharan, et al. (2017)	✓			✓					✓		✓	✓			✓	UG	B & C	PL
Raman et al. (2013)	✓			✓					✓		✓	✓			✓	UG	B & C	PL
Xu et al. (2020)	✓					✓			✓		✓		✓			UG	A & C	P
J. Kim and Canny (2017)	✓		✓			✓			✓		✓		✓	✓		UG	B & A	P
Cultrera et al. (2020)	✓			✓					✓		✓		✓	✓		UG	B & A	P
Schneider et al. (2021)	✓			✓					✓				✓	✓		ED	A & C	P
Rahimpour et al. (2019)				✓					✓		✓		✓	✓		UG	B & C	P
Shen et al. (2020)	✓					✓			✓		✓		✓	✓		ED	B	P
Ben-Younes et al. (2020)	✓					✓			✓		✓		✓	✓		UG	A & C	P
Nahata et al. (2021)	✓		✓	✓				✓		✓			✓	✓		UG	B	PL
Ha et al. (2020)	✓			✓					✓			✓				ED,PC	A & B	P
Koo et al. (2015)	✓			✓					✓			✓				ED,PC	A & B	P
Bojarski et al. (2018)	✓					✓			✓		✓		✓	✓		UG	B & C	P
Mori et al. (2019)	✓			✓					✓		✓		✓	✓		UG	B & C	P
T. Liu et al. (2018)	✓			✓					✓			✓		✓		ED	A	P
Omeiza, Kollnig, et al. (2021)	✓	✓	✓	✓					✓			✓				ED	A	P
Rizzo et al. (2019)	✓			✓					✓		✓		✓	✓		UG	B	P
Y.-C. Liu et al. (2020)	✓			✓					✓		✓		✓	✓		UG	B & C	P
Omeiza, Web, et al. (2021)	✓	✓	✓	✓					✓			✓				ED	A	P

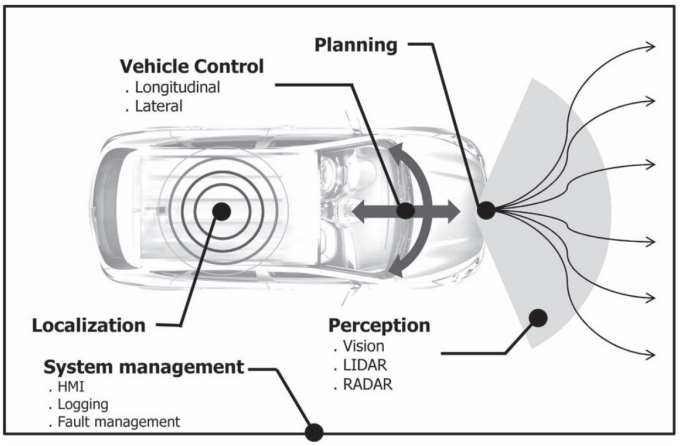


Figure 2.2: Key operations of an autonomous vehicle (Jo et al., 2014). In Section 2.5 and Section 2.6, we discuss the role of explanations within these key operations.

Figure 2.2 illustrates the different AV driving operations.

2.5.1 Perception

Autonomous vehicles rely on cameras placed on every side—front, rear, left and right—to stitch together a 360-degree view of their environment. These cameras range from a wide field of view and a shorter range, to a narrow view for long range visuals. Though they provide accurate visuals, cameras have their limitations. They can distinguish details of the surrounding environment, however, the distances of those objects need to be calculated to know exactly where they are. It is also more difficult for camera-based sensors to detect objects in low visibility conditions, like fog, rain or nighttime (Zang et al., 2019).

Radar sensors can supplement camera vision in times of low visibility, like night driving, and improve detection for self-driving cars. These radar sensors work by transmitting radio waves in pulses. Once those waves hit an object, they return to the sensor, providing data on the speed and location of the object. While the data provided by surround radar and camera are sufficient for lower levels of autonomy, they are unable to cover all situations without a human driver (Piramuthu & Caesar, 2021).

Lidar is a sensor that measures distances by pulsing lasers which makes it possible for self-driving cars to have a 3D view of their environment. It provides

shape and depth to surrounding cars and pedestrians as well as the road geography. Similar to the radar, the Lidar sensor is able to function in low-light conditions. By emitting invisible lasers at incredibly fast speeds, Lidar sensors are able to paint a detailed 3D picture from the signals that bounce back instantaneously. These signals create "point clouds" that represent the vehicle's surrounding environment to enhance the safety and diversity of sensor data (Y. Li & Ibanez-Guzman, 2020).

Camera, radar and lidar sensors provide rich data about the car's environment. The different inputs from the sensors are sometimes fused and processed by deep learning algorithms (black-box models) for scene understanding tasks. Hence, the need for datasets for the design of explainable driving models.

Driving Datasets For Posthoc Explanations

Several driving datasets have been made available for the purpose of training machine learning models for autonomous vehicles (Janai et al., 2020). Some of these datasets have annotations—e.g., handcrafted explanations (J. Kim et al., 2018; You & Han, 2020), vehicle trajectories (Houston et al., 2020), human driver behaviour (Ramanishka et al., 2018; Shen et al., 2020) or anomaly identification with bounding boxes (Xu et al., 2020; You & Han, 2020)—that are helpful for posthoc driving behaviour explanation. We have categorised the sensors used in collecting the datasets into exteroception and proprioception types, and the annotations in the datasets that are useful for developing explainable AVs. We also identified different stakeholders that can potentially benefit from the explanations. See Table 5.4). Although the datasets are helpful for developing explanation methods, it is important to note the potential challenges associated with the use of these datasets. Each dataset was collected from one region of the world, thus, chances are high that they may not generalise, especially where traffic signs, rules, and road topology are quite different to that of other regions; this can potentially lead to biased driving decisions. Also, most of the datasets only provide a video of the external environment and do not provide internal AV state data. It is therefore a concern as to whether the explanation techniques designed with this dataset will be very faithful to the AV.

Table 2.4: Driving datasets that are useful for developing explanation methods for AVs and the stakeholders that would potentially benefit from such explanations. **Ext.**: Exteroception, **Prop.**: Proprioception.

Dataset	Size	Ext. Cam	Prop. CAN	Annotation & Explanation	Stakeholders (see Sec. 2.3.3)
BDD-X (J. Kim et al., 2018)	7K × 40s	✓	✗	Textual *Why explanation* associated with videos segments with heatmaps	Class A, B, and C
BDD-OIA (Xu et al., 2020)	23K × 5s	✓	✗	Actions and *Why explanation*	Class A, B and C
DoTA (Yao et al., 2020)	4,677 videos (73,193s)	✓	✗	*What explanation* (Temporal and spatial anomaly identification with bounding boxes)	Class B and C
CTA (You & Han, 2020)	1,935 × 17.7s	✓	✗	Why explanation for accidents with cause and effects	Class B and C
HDD (Ramanishka et al., 2018)	374,400s	✓	✓	*What explanations* for driver actions	Class B
BDD-A Extended (Shen et al., 2020)	1,103 × 10s	✓	✗	Human gaze inciting *why and/or what explanation*, explanation necessity score	Class B
Lyft Level5 (Houston et al., 2020)	360,000s	✓	✗	Trajectory annotation	Class B

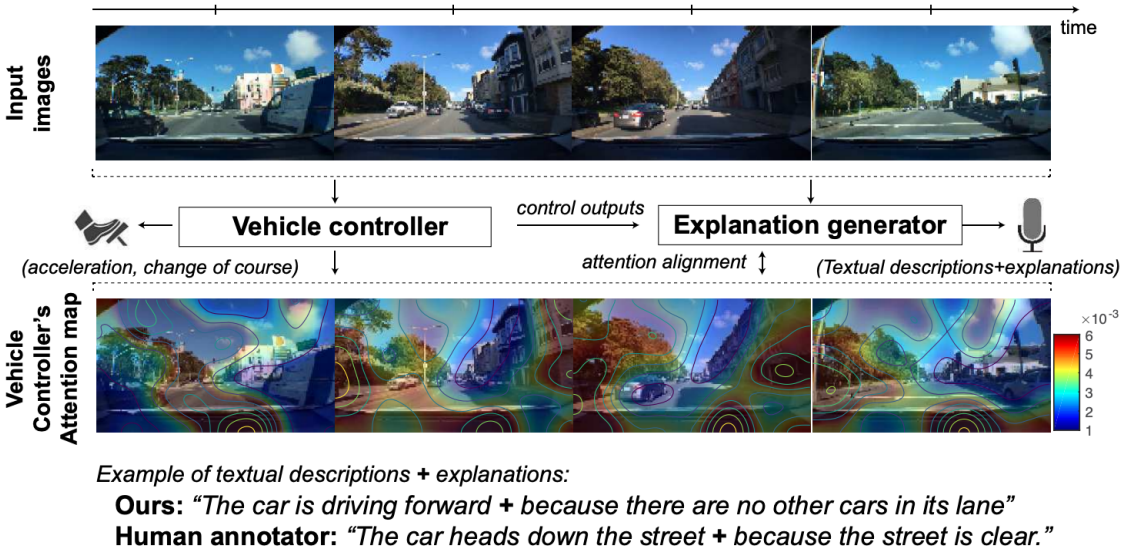


Figure 2.3: The vehicle control model predicts commands such as acceleration and a change of course, the explanation generator model generates natural language explanations and attention maps (J. Kim et al., 2018)

.

Vision-Based Explanations for AVs

Various methods have been proposed to explain neural networks which are fundamental structures for perception and scene understanding in AVs. Some of the prominent methods are *gradient-based*. Gradient-based or backpropagation methods are generally used for explaining convolutional neural network models. The main logic of these methods is dependent on gradients that are backpropagated from the output prediction layer of the CNN back to the input layer (Das & Rad, 2020). They are often presented in form of heatmaps (see Figure 2.3). These methods mainly fall under the input influence explanation style in the explanation categorisation presented in Table 2.3.

I provide some examples of gradient-based methods that are useful for explanations in AV perception. Refer to (Tjoa & Guan, 2019; Zablocki et al., 2021) for a survey on vision-based explanation methods.

- Class Activation Map (CAM) (B. Zhou et al., 2016) and its variants like Gradient Class Activation Map (Grad-CAM) (Selvaraju et al., 2017), Guided Grad-CAM (Tang et al., 2019), Grad-CAM++ (Chattopadhay et al., 2018), Smooth Grad-CAM++ (Omeiza et al., 2019)

- Other gradient-based methods include VisualBackProp (Bojarski et al., 2018), Layer-wise Relevance Propagation (LRP) (Lapuschkin et al., 2019; Samek et al., 2017a), DeepLift (Shrikumar et al., 2017), (Zeiler & Fergus, 2014), and Guided-Backpropagation (Springenberg et al., 2014).

Many of the vision-based explanations for AVs stem from the generic gradient-based methods explained above. For example, Bojarski et al. (2018) proposed VisualBackProp for visualising super-pixels of an input image that is most influential to the predictions made by a CNN model. In Bojarski et al. (2018), VisualBackProp on an end-to-end learning model for autonomous driving—PilotNet (Bojarski et al., 2016)—was applied to check whether the explanation method is able to show the parts of a driving scene image that are necessary for the steering operation of the AV model.

J. Kim et al. (2018) proposed an approach for explanation generation in autonomous driving. The approach involves training a convolutional neural network end-to-end from images to the vehicle control commands (which are acceleration and change of course). Further, textual explanations of the model actions are produced through an attention-based video-to-text model trained on the BDD-X dataset. Explanations were provided in form of saliency maps and text (see Figure 2.3). A related work by Xu et al. (2020) focused on scene understanding, highlighting salient objects in input that can potentially lead to a hazard. These objects are described as action-inducing since their state can influence the vehicles' decisions. Apart from identifying objects, a sequence of short explanations was generated.

2.5.2 Localisation

Localisation in AVs is the process of determining the pose (e.g., location and orientation) of the AV relative to a piece of given information (e.g., map) of the environment. Precise and robust localisation is critical for AVs in complex environments and scenarios (L. Wang et al., 2017). For effective planning and decision-making, the position and orientation information is required to be precise in all weather and traffic conditions. One of the goals of a precise and robust localisation is to ensure that the AV is aware of whether it is within its lane (Reid et al., 2019) for safety purposes. Safety is often considered the most important design requirement and it is critical in the derivation of requirements for AVs (Reid et al., 2019). Hence, communicating position over time and with justifications as explanations is crucial to expose increasing error rates in a timely manner before they cause an accident. For instance, the position errors can be transmitted continuously through a wireless channel to an operation centre from which the AV is managed. An interface that displays this information (e.g., a special dashboard or mobile application as shown in (Schneider et al., 2021) is provided and it is able to trigger an alarm for immediate action (e.g., safe parking) when the error margin is exceeded.

Although there seems to be less research related to explainable localisation, intelligible explanations remain key. They would allow for easy communication

of the position of an AV, including the measurement's precision and error (Reid et al., 2019), of an autonomous vehicle during the localisation process in the form of clear and intelligible explanations. Explanations from localisation will be handy for Class B stakeholders (i.e., system developers) for debugging AVs because it can facilitate positional error correction and provide other stakeholders perception of reliability and safety for AVs. Potentially, it will inform the development process of more robust localisation procedures

2.5.3 Planning

Through AI planning and scheduling, the sequence of actions required for an agent to complete a task is generated. These action sequences are further utilised in influencing the agent's online decisions or behaviours with respect to the dynamics of the environment it operates in (Ingrand & Ghallab, 2017). The planning system is an important aspect of autonomous vehicles because of the complex manoeuvres they make in dynamic, complex, and sometimes less structured or cluttered environments (e.g., urban roads, street roads with lots of pedestrians and other road participants). In fact, traffic elements (e.g., roadside infrastructures, road networks, road signs, and road quality) are dynamic and can change with time; this makes AVs regularly update their plans (and even learn sometimes) as they operate. Often, the amount of data (e.g., descriptions of objects, states, and locations) that the AV processes per time is larger than such that a human may be able to process, and continuously and accurately keep track of. Hence, a stakeholder riding in an AV may be left in a confused state when the AV updates its trajectory without providing an explanation.

Explainable planning can play a vital role in supporting users and improving their experiences when they interact with autonomous systems in complex decision-making procedures (Chakraborti et al., 2020a). According to (Sado et al., 2020), depending on the stakeholder, the process may involve the translation of the agent's plans into easily understandable forms, and the design of the user interfaces that facilitate this understanding. Relevant work include XAI-PLAN (Borgo et al., 2018), WHY-PLAN (Korpan & Epstein, 2018), refinement-based planning (RBP)

(Bidot et al., 2010), plan explicability and predictability (Zhang et al., 2017), and plan explanation for model reconciliation (Chakraborti, Kulkarni, et al., 2019; Chakraborti et al., 2020b; Chakraborti, Sreedharan, et al., 2017).

XAI-PLAN is a domain-independent, planning system agnostic, and explainable plan model that provides initial explanations for the decisions made by an agent planner (Borgo et al., 2018). The user explores alternative actions in a plan and a comparison is done with the user's resulting plan and the plan that was suggested by the planner. The XAI-PLAN framework then provides an explanation to justify discrepancies. This kind of interaction encourages and enhances mixed-initiative planning which has the potential to improve the final plan. Interestingly, users can pose contrastive forms of queries in the form "why does the plan contain action X rather than action Y?".

Refinement-based Planning (RBP) A related transparent and domain-independent framework called refinement-based planning (RBP) (Bidot et al., 2010) produces explanations of verbal plans upon a verbal query from a user. It possesses an enhanced representation of the search space, providing a 2-way search (i.e., forward and backwards) capability when generating plans. This allows for flaw detection and plans update or optimization. Using states and action primitives, the RBP paradigm integrates partial-order causal-link planning and hierarchical planning (Biundo & Schattenberg, 2014) (hybrid planning framework).

Why-Plan Korpan and Epstein (2018) also proposed Why-Plan, an explanation technique in human-machine collaborative planning. The method juxtaposes a person's and an autonomous agent's objectives in a path planning process and provides explanations to justify the differences in planning objectives in a meaningful and human-friendly fashion. It basically addresses questions like "why does your plan involve that action?"

The explainable planning frameworks described above and the related work by (Chakraborti, Sreedharan, et al., 2019; Chakraborti, Sreedharan, et al., 2017;

Hayes & Shah, 2017; Neerincx et al., 2018) can serve as basics to build upon for plan explanations in AVs.

2.5.4 Vehicle Control

Control in an AV generally has to do with the manipulations of vehicle motions such as lane changing, lane-keeping, and car following. These manipulations are broadly categorised under longitudinal control (speed regulation with throttle and brake) and lateral control (i.e., automatic steering to follow track reference) (Khodayari et al., 2010).

ADAS currently works based on the AV's sensor information obtained from observing the environment. Interfaces that come with ADAS now display rich digital maps ("TomTom launches map-based ADAS software platform Virtual Horizon", n.d.), vehicle's position, and track related attributes ahead or around the vehicle. Stakeholders may issue investigatory queries when the AV makes a decision against their expectations. For instance, the stakeholder may want to ask different questions based on current contexts (e.g., near-miss, special vehicle case, or collision). Investigatory queries could be in form of a 'why' question (e.g., 'Why did you turn left?'), 'why not' or contrastive question (e.g., 'why did you switch to the left lane instead of the right lane'), 'what if' or counterfactual questions (e.g., 'what if you turned left instead of right?'), or 'what' question (e.g., 'What are you doing?').

Other than existing in-vehicle visual interfaces such as mixed reality (MR) visualization (Sasai et al., 2015), and other flexible (i.e., highly reconfigurable) dashboard panels (Marques et al., 2011), in-vehicle interfaces that support the exchange of messages between the stakeholder and the AV is crucial. The user should be able to query the interface and receive explanations for navigation and control decisions in an appropriate form; either through voice, text, visual, gesture or a combination of any of these options.

In the next section, we review explanations in relation to AV system management and interaction with respective stakeholders.

2.6 AV System Management

In this section, we review works relating to event data recording (EDR) in AVs and human-machine interactions involving in-vehicle interfaces and external human-machine interfaces (eHMI) that could be potentially used for explanations.

2.6.1 Logging and Fault Management: Event Data Recorder

The event data recorder (EDR) serves as a recording device in automobiles to log information related to vehicle accidents. Upon a posthoc analysis, a better understanding of how certain faults or accidents come about is achieved (Wu et al., 2013).

The installation of EDR in passenger vehicles has been a mandatory process in the United States since 2014. Recently, the National Transportation Safety Board (NTSB) suggested the need for risk mitigation pertaining to monitoring driver engagement and the need for better event data recording requirements for autonomous driving systems after the Tesla crash case in 2018 (NTBS, 2018).

As autonomous vehicles increase in society and gain more public attention, it is necessary to discriminate human driver errors and negligence from the AV's errors—arising from non-adapted or poor product design or a product defect (Bose, 2014; Kohler & Colbert-Taylor, 2014)—and express these errors in explanations. Martinesco et al. (2019) attributed the existing challenge—in ascribing faults to the appropriate traffic participant—to the difficulty in identifying and evaluating the correct cause of an accident.

In line with this, the National Highway Traffic Safety Administration (NHTSA) calls for the industry and standard bodies such as SAE and IEEE to develop a uniform approach to address data recording and sharing (see relevant document (NHTSA, n.d.) which may, in turn, be useful for explanations. Pinter et al. (2020) deplored the inability of the existing EDRs to provide sufficient data needed to reconstruct the behaviour of a vehicle before and after an accident, and to a degree that the accident could be analysed from the perspective of liability. As AV functions continue to increase (eventually leading to full autonomy), the storage of a satisfactory number of parameters is needed for the reconstruction of the

vehicle's behaviour and the provision of explanations for a reasonable amount of time before and after the accident becomes crucial.

As an effort towards building more effective EDRs that can support explanation provision, different approaches, which include the use of blockchain technologies, and more effective and robust data models have been proposed. Guo et al. (2018) proposed a blockchain-inspired EDR system for autonomous vehicles to achieve indisputable accident forensics by providing trustability and verifiability assurance of an event's information. With this blockchain approach, the verification and confirmation of a new block of event data are possible with no central authority involved. In terms of storage mechanisms and reliability, Yao and Atkins (2020) proposed a Smart Black Box (SBB) to supplement traditional data recording with value-driven higher-bandwidth data capture. The SBB uses a deterministic mealy machine (Harris & Harris, 2013) based on data value and similarity to cache short-term histories of data as buffers. By optimising value and storage cost trade-offs, the appropriate compression quality for each data chunk in the driving history data is determined. Prioritised data recording prevents the retention of low-value buffers. By discarding them, space is made available to store new data.

With the EU legislative rules on EDR enacted in 2022 ("TEuropean Commission - Press release: Road safety: Commission welcomes agreement on new EU rules to help save lives", 2019)—and a similar one in China (UNECE, 2019)—there is the question as to whether existing data storage facilities are sufficient for the data needs for accident investigations involving automated vehicles. For efficient storage space management, a well-defined data package which puts the data points (with necessary parameters) and the frequency of measuring and recording that can enable full reconstruction of AVs' regular and irregular movements is necessary for event explanation purposes. The data model from Pinter et al. (2020) can be used to determine the data content required in an EDR, sufficient for accident investigations, and suitable for vehicles at different autonomy levels. Further, Böhm et al. (2020) proposed a broader database in relation to the US EDR regulation (NHTSA 49 CFR Part 563.7) after carrying out a study involving the reconstruction of real accidents

with ADAS enabled vehicles to investigate requirements. These advancements in EDRs are relevant for the development of explanation techniques for accidents (and other critical events). It may also draw researchers' attention to explainable EDR which is currently very much under-explored. Human-machine interaction (HMI) is a key aspect of explanation in AVs. In the next section, we will discuss the relationship between HMI and explanations in the autonomous driving context.

2.6.2 Human-Machine Interaction

Human-Machine Interaction can be viewed from two different perspectives in automated driving. First, interaction between two or more road users (e.g., AV and pedestrian). This is a situation where the behaviour of at least two or more road users can be interpreted as being influenced by the possibility that they are both intending to occupy the same region of space at the same time in the near future (Markkula et al., 2020). Secondly, the interaction between an in-vehicle participant and the in-vehicle interfaces. We examine related works across these perspectives.

Generally, AVs possess components for sensing, decision-making, and the operation of the vehicles, requiring minimal human driving (Smith & Svensson, 2015). They can operate in complex environments where the decision set is large (Yurtsever et al., 2020b). This poses a challenge to the understandability of their operational modes. Vehicles are seen to have evolved over the years in terms of automation level, and in-vehicle technologies and interfaces (i.e., technologies and interfaces within the vehicle). Essentially, vehicles in the SAE levels 0 to 2 have a low explanation requirement due to their low complexity. For vehicles in levels 3 and above, the explanation requirement is high due to their high complexity. Table 2.5 provides a summary based on (Edwards, 2014; Lavrinc, 2018), with SAE levels and explanation requirements. As shown in Table 2.5, vehicle instrument (i.e., an instrument that measures some quantities about the vehicle) interfaces evolved to adaptive displays where content is presented in a form that enhances user experience, and with enhanced positioning features, e.g., global navigation satellite system (GNSS).

Table 2.5: Vehicle Instrument Interface Evolution and Explanation Need

#	Explanation Interface	SAE Automation Level	XAI Demand	Vehicle Examples
1	Fully analogue interface	Level 0	Low	Old Ford vehicles and similar vehicles back before the year 1990
2	Partly analogue and digital interface (e.g., digital odometer, analogue speed dial)	Level 0	Low	Older Honda Civics, Citroen C4 Picasso and others mostly between 1990 and 2000.
3	Mostly digital interface	Level 0 and 1	Low	BMW 5 Series, Fiat 500, and Jaguar XF and others mostly between 2010 and 2016
4	Fully digital interface with adaptive display, GNSS)	Level 2 and 3	Moderate	Tesla Autopilot, Audi A8 2016 to present
5	Fully digital interface with adaptive display, Sat Nav)	Level 4	High	Waymo cars 2016 to present

Recent highly automated vehicles are incorporating more enhanced interaction technologies. Moreover, novel interaction technologies provide the opportunity for the design of useful and attractive in-vehicle user interfaces that abstract and explain vehicle automation operations (e.g., perception, planning, localisation, and control) exist. In the next section, we will discuss previous research on in-vehicle interfaces.

Novel Interaction Technologies

The in-vehicle user interface is essential for efficient explanation provision, and in enhancing driving experience (Schmidt et al., 2010). There are studies that suggest that interface design trends impact driving experience. For example, Jung et al. (2015) explored the impact of the displayed precision of instrumentation estimates of range and battery state-of-charge on drivers' driving experience, and attitude towards varying conditions of resource availability in an all-electric vehicle. Results from the study showed that it can be advantageous to display the uncertainty values associated with a measure rather than concealing it as participants presented with an ambiguous display of range measure reported a preserved trust level towards the vehicle. Although presenting users with a single number value increased reading and apprehension time, the implication of disguised uncertainty on user experience and behaviour has to be carefully considered in critical situations.

A related work by Mashko et al. (2016) involved the assessment of in-vehicle navigation systems with a visual display where virtual traffic signs were represented

on an in-vehicle display to assist better orient at road sections loaded with excess information clutter. The use of virtual traffic signs in-vehicle improved the drivers' concentration and reaction to traffic signs on the road. Langlois (2013) proposed an interface (Lighting Peripheral Display—LPD) that creates signals that are able to be handled by peripheral vision (the ability to see objects and movement outside of the direct line of vision) while driving in order to enhance the utility of ADAS. The LPD possessed a box illuminated by light-emitting diodes (LEDs) and reflected onto the windscreen. User tests conducted showed that driving performance and comfort were enhanced by LPDs. Sirkin et al. (2017) developed Daze, a technique for measuring situation awareness through real-time, in-situ event alerts. The technique is ecologically valid in that it is very similar in look and feel to the applications used by people in actual driving environments, and can be applied in simulators and also in on-road research settings. The authors conducted a study that included simulated-based and on-road test deployments in order to provide assurance that Daze could characterise drivers' awareness of their immediate environment and also understand the practicality of its use.

Having examined the existing interaction technologies, it would be worthwhile to look at what users actually prefer.

In-vehicle Interfaces—User Preferences

Learning about the experiences of in-vehicle participants will help to inform what users' preferences are. Mok, Sirkin, et al. (2015) described a Wizard of Oz study to get insights into how automated vehicles ought to interact with human drivers. Design improvisation sessions were conducted inside a driving simulator with interaction and interface design experts. While the two human operators (wizards) controlled the audio and driving behaviour of the car, the participants were driven through a simulated track with different terrain and road conditions. The study noted that:

1. instead of taking over full control, participants wanted to share control with the vehicle;

2. participants like to know exactly when a handover (mode switch) happens and require a clear alert from the vehicle to that effect;

3. to the participants, delayed responses and unperformed requests were acceptable as long as the responses provided are correct/proper;

4. AVs have a variety of means to help sustain or improve participants' trust in them.

In a related study highlighting the significance of in-vehicle alert systems through HMIs, Fu et al. (2020) examined the effect of varying sensitivity and automation levels of in-vehicle collision avoidance systems. This was done using the automatic emergency braking (AEB) systems in Level 3 vehicles where a driver is needed to monitor the system for failures. Drivers reacted more (in terms of vigilance and awareness) to the system when it was biased to under-report hazards. The result also suggested that higher levels of automation result in lesser driver vigilance and awareness, resulting in significantly worse driver performance. A similar study examined how drivers would behave when they are subjected to an unstructured emergency transition of control in the presence of an audible alert (Mok, Johns, et al., 2015).

Regarding consumer preferences, Park et al. (2020) conducted a study in an attempt to understand the extent to which semi-AV decision-making should account for individual user preferences. Having considered 18 different scenarios with tactical driving goals, significant differences were discovered in scenario interpretations, AV perceptions, and vehicle decision preferences. The alignment of individual preference with AV decision yielded more positive changes in the consumer impression of the vehicle than unaligned decisions.

As more AVs use deep vision-based approaches for scene understanding, the probabilistic nature of these approaches introduces varying degrees of uncertainty in object detection and scene understanding, which are essential for path planning. Communicating these uncertainties to drivers or operators is critical for safety reasons. A. 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, A. Kunze et al. (2019b) used peripheral awareness display to communicate uncertainties with the aim of alleviating 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, Eder, et al., 2021), scene detection and prediction (Colley et al., 2022), pedestrian detection and prediction (Colley et al., 2020). 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.

Closely related to natural explanations, Ha et al. (2020), Koo et al. (2015), and Omeiza, Kollnig, et al. (2021) investigated the effect of explanations on trust through empirical user studies. Ha et al. (2020) examined two explanation types, simple and attributional, as well as perceived risk on trust in AVs in four autonomous driving scenarios with varying levels of risk using a simulation of an in-vehicle experience. Their results indicated that an explanation type can greatly affect trust in autonomous vehicles and that under high levels of perceived risk, attributional explanations lead to the highest trust in AVs.

Further, Schneider et al. (2021) investigated whether the provision of explanations in simulated driving can enhance user experience and increase the subjective feeling of safety and control. The provision of explanations did not influence user experience during and after the ride. See an example of a basic mobile explanation interface in Figure 2.4.

These works highlight the importance of the user-centred approach (which could have a great influence on the future of human-machine interaction) in the design and development of explanation methods for AVs. In-vehicle user interfaces are a medium for the provision of visual, text, or voice explanations. However,

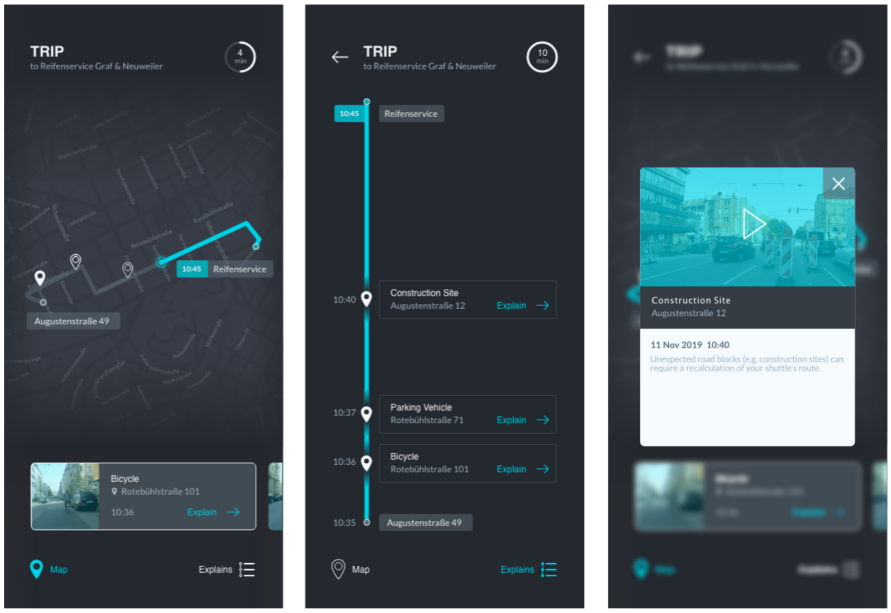


Figure 2.4: An example of a mobile interface for an AV explainer (Schneider et al., 2021). This interface is used for posthoc explanation provision. The app provides a record of a journey and can provide explanations from three different views at strategic points of the journey.

previous works only focus on providing information about the vehicle to users without communicating reasons and/or causal links for decisions. This remains an open challenge for AV designers and researchers. I now explore the interaction between an AV and other external traffic participants.

AV and External Agents Interaction

There are different categories of traffic participants that an AV has to interact with, and there are many studies focusing on the interactions between AVs and other traffic participants (Eby et al., 2016; Yang & Coughlin, 2014). The traffic participants that AVs will frequently interact with are pedestrians, cyclists and other vehicles.

Pedestrians Pedestrians and human drivers communicate intents to each other to inform the next choice of action (Šucha, 2014; Sun et al., 2003; Wilde, 1980). Studies on driver and pedestrians communication strategies (Clamann et al., 2017; Y. M. Lee et al., 2021) suggest that explicit eHMIs are less likely to be used by pedestrians during road crossing compared to vehicle-based movement information such as

yielding cues. However, it has been argued that it is important that autonomous vehicles need to have modalities for communicating intents to pedestrians (Lagstrom & Lundgren, 2015; M. Faas et al., 2021; Rasouli & Tsotsos, 2019).

Mahadevan et al. (2018) conducted a study to get insights into interface designs that explicitly communicate autonomous vehicle awareness and intent to pedestrians. Different interface prototypes were developed and deployed in a study that involved a Segway and a car in a simulation setting. Results suggest that interfaces communicating vehicle awareness and intent can assist pedestrians attempting to cross at crosswalks and can exist in the environment outside of the vehicle. They suggested a combination of modalities (e.g., visual, auditory, and physical) in the interfaces.

M. Faas et al. (2021) investigated pedestrians' trust and crossing behaviour in repeated encounters with AVs in a video-based laboratory study. The occurrence of AV malfunction and system transparency with status and intent eHMI were studied. Their results showed that trust increases with the presence of status and intent eHMI and decreases when there is a malfunction in the AV but recovers quickly. Crossing onset time also decreased with the provision of the eHMI. Crossing onset time indicates the time in seconds between the vehicle yielding and the pedestrian stepping off the sidewalk (Faas et al., 2020). It was noted that status eHMI can cause pedestrians to overtrust AVs, therefore, intent messages are needed to complement status eHMIs.

Pedestrians' reactions to a ghost driver have also been investigated in the academic literature. A ghost driver in this context refers to a driver who pretends to be absent in the car even when they are in control of the driving operations. Moore, Currano, et al. (2019) conducted a Wizard-of-Oz driverless vehicle study aimed to test pedestrians' reactions to everyday traffic in the absence of an explicit eHMI. Although some pedestrians were surprised by the vehicle's supposed autonomy, others neither noticed nor paid attention to its autonomous nature. All the pedestrians crossed in front of the vehicle without explicit signalling. This suggests that the vehicle's implicit eHMI (which is basically its observed motion) may

suffice. Therefore, pedestrians may not need the explicit eHMI in their interaction routine. A similar study by Moore et al. (Moore, Strack, et al., 2019) indicated that pedestrians crossed in front of a ghost vehicle with little hesitation even when the vehicle did not give any signal beyond its motion. However, J. Li et al. (2020) findings contradict this claim by confirming pedestrians' behaviours are different on encountering a vehicle with a hidden driver based on a study carried out in Europe.

Interaction with Pedestrians with Reduced Mobility (PRM) Pedestrians with reduced mobility might need their support devices re-engineered to allow for effective interaction with AVs (Asha et al., 2020). Asha et al. (2020) carried out a design study to explore interface designs for interaction between AVs and PRMs. The results from the analysis disclosed that visual cues are the most important interface elements, and street infrastructures are the most important location for housing cues for this category of pedestrians. They also found that wheelchairs might require an interface, and the current wheelchairs would have to be altered to allow for this interface.

Other Road-Participants Vehicle-cyclist interaction is an important topic to examine, especially in an environment where cycling is common. Cyclists and drivers currently communicate through implicit cues (vehicle motion) and explicit but imprecise signals such as horns, lights, and hand gestures (Hou et al., 2020). Virtual reality (VR) AV-cyclist immersive simulators and a number of AV-cyclist interfaces have been designed to study interactions between AVs and cyclists. Findings from one of such studies (Hou et al., 2020) suggest that AV-cyclist interfaces can improve rider confidence in lane merging scenarios. Future AVs could consistently communicate feedback (in form of explanations) to create awareness and indicate intents, leveraging their sensor data.

In general, more research is needed to explore how the findings from these studies can be utilised to create effective and efficient interaction interfaces between AVs and stakeholders in order to facilitate the provision of explanations.

In the next section, we will present some challenges around explainability in autonomous driving and consider them as open research questions which guide the rest of the research in this thesis.

2.7 Research Gaps

I highlight three limitations and the corresponding research opportunities motivating the research presented in the subsequent chapters of this thesis.

2.7.1 Human Factors

Limited User-Centric Explanation Design and Assessment

As seen in Table 2.3, research on explanations in AVs has mainly focused on the theory and implementation of explanations based on perception data with less user-centric empirical studies. There is a scarcity of rigorous user studies to elicit stakeholders' explanation requirements which include when an explanation is needed and the appropriate type for each scenario and stakeholder, especially those in Class A. Moreover, the previous human-centric studies on explanations in autonomous driving have used hand-crafted explanations. Manual generation of natural language explanations is not possible at the scale required for AV deployment. Thus, there is a need for efficient and robust automatic generation of useful natural language explanations in the autonomous driving domain. Further, existing research has not intentionally and adequately explored explanation related theories from behavioural sciences. For example, the theory of causal attribution, contrastiveness of explanations, counterfactual reasoning, and folk psychology from the behavioural sciences are important concepts to consider in providing explanations (Miller, 2019), especially in autonomous driving. Hence, we build atop some of these theories in the following chapters.

2.7.2 Technical Factors

Implementation Limitation: Transparency and Faithfulness

Previous works have provided natural language explanations to accompany attention maps (S. Chen et al., 2021; J. Kim et al., 2018). However, the natural language text generation techniques usually applied in these works involved the use of a complex language model to learn relationships between extracted features and handcrafted textual explanations. This poses the 'chicken and egg problem'—explaining a deep learning vision-based model with a complex language model. Moreover, as the ground truth textual explanations were provided by a third party who was independent of the driving process, the faithfulness of the resulting explanation to the driving model is questionable. Furthermore, there have been no considerations for internal state data in the development of the previous explanation methods. To improve the faithfulness of explanations, apart from the perception data, internal state data which can be obtained from an AV's CAN bus should be leveraged. Transparency—which could be achieved by the use of interpretable/transparent approaches—and faithfulness are open research opportunities that we have explored in this thesis.

Presentation Limitation: Intelligibility

As elucidated in the survey, there are a couple of research works in which explainable AI algorithms have been applied to explain deep models for driving tasks. These methods are based on saliency in the sense that the resulting explanations are heatmaps indicating the important pixels/portion in the image for predicting a class. Examples are CAM (B. Zhou et al., 2016), Grad-CAM (Selvaraju et al., 2017), and attention weights heat maps (visual attention maps) as used in (J. Kim et al., 2018). These saliency methods create spurious heat maps, with high entropy or noise. Thus, not useful in easily understanding the behaviour of the underlying model of the system. This problem is further elaborated in a recent seminal work on sanity checks for saliency methods (Adebayo et al., 2018). Moreover, where heatmaps are not used—e.g., in (Chakraborti, Kambhampati, et al., 2017) where

textual explanations of plans were generated—the explanations are yet too technical and are not communicated in natural language. This is inappropriate for lay users. Hence, there is a research opportunity to develop explanation techniques that would generate intelligible natural language explanations. As a start, high-level commands (e.g., turning right, lane change left, and lead vehicle accelerating, among others) as used in the National Highway Traffic Safety Administration (NHTSA) report (Najm et al., 2007) can be used to represent transitions between road and lane segments, and interaction with other road participants.

2.7.3 Regulatory Factors

Standards and Regulations

Some of the standards provided in Table 2.1 are very relevant to explainability in autonomous driving. For example, ISO/TR 21707:2008, which specifies a set of standard terminology for defining the quality of data exchanged between data suppliers and data consumers in the ITS domain, is very relevant to AV explainability, although not originally intended for explainability. While data quality is important, the presentation style, language, and the interfaces by which the data is provided are also critical for explanations in autonomous driving. We suggest that this standard and others in Table 2.1 be explored for the development of more AV explainability related ones and should be made easily accessible.

Regulations regarding the explainability of automated systems are being set by countries and regions. However, these regulations seem quite abstract and do not directly address requirements in line with AV technologies and the stakeholders involved. For example, in the preliminary consultation paper on autonomous vehicles (Law Commission, n.d.), the UK Law Commission states the recommendation of the National Physical Laboratory on explainability in autonomous driving as follows:

> It is recommended that autonomous decision-making systems should be available, and able, to be interrogated post-incident. Similar to GDPR, decisions by automated systems must be explainable and key data streams stored in the run-up, during and after an accident.

There is no information on the nature of the explanation and level of detail to provide the different AV stakeholders, as explanation requirements differ across stakeholders. Moreover, specifics as per the explainability requirements for each component of the AV stack are missing. This makes the realisation of explainable AVs challenging.

In this thesis, we mainly focus on the human and technical factors as evidenced by the research questions, but offer regulatory recommendations in Chapter 7, the concluding chapter. In the next chapter, we will address the first research question by exploring different explanations and driving scenarios, and then describe a user study conducted to provide a better understanding of end-users' preferences.

3

Explanation Requirements in AVs: An Empirical Study

Contents

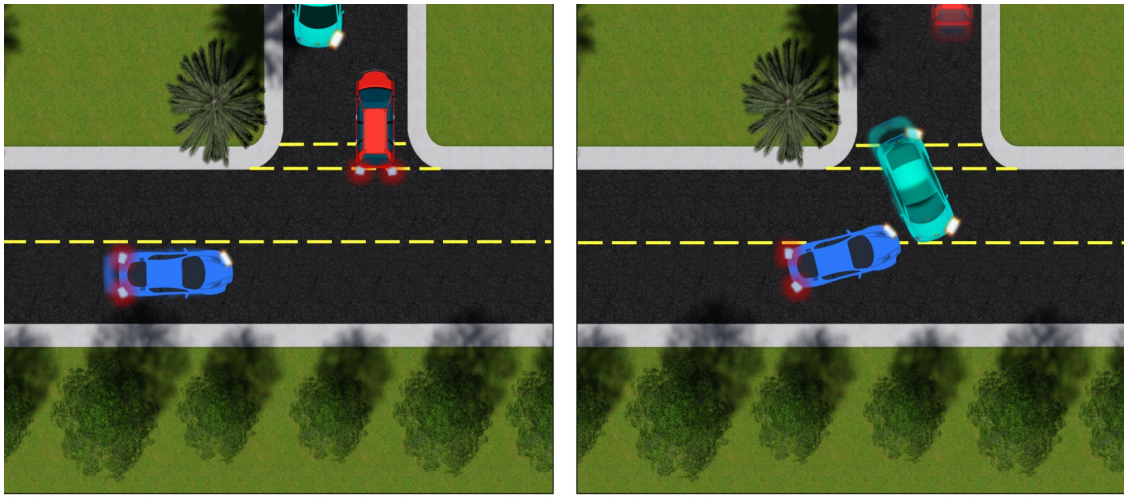


Figure 3.1: The scenario depicts a *near miss* situation between an autonomous vehicle (AV) (blue) and another traffic participant (green). The green vehicle failed to yield, but the AV adjusted to avoid a collision. Here are two out of the four types of explanations provided: **Non-Causal Explanation (*What*):** 'We stopped to avoid a collision with the green vehicle.' **Causal Explanation (*Why Not*):** '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.'

3.1 Introduction

In this chapter, we address the first research question on the investigation of different explanations under different driving scenarios. We describe a between-subject user study carried out to investigate the effect of different explanation types provided by an AV to in-vehicle participants (passengers specifically) in a range of driving scenarios. Participants were asked to engage with groups of image sequences illustrating different driving scenarios with corresponding explanations (see sample scene in Figure 3.1). We assessed participants' understanding of the driving situations (intelligibility), accountability, and their perceptions of trust. We categorised explanations based on the causal filters (i.e., the *Why*, *Why-Not*, *What-If*, and *What*). Generally, explanations that explicitly state reasons for an effect can be triggered by the *Why*, *Why-Not*, *What-If* filters; we refer to this class of explanations as explanations with causal attribution or simply causal

explanation (Kelley, 1973). The *What* filter is generally not expected to trigger explanations that provide reasons for an action. It only provides state information; we refer to explanations from the *What* filter as explanations without causal attribution or non-causal explanation. See Table 2.2.

3.1.1 Hypotheses

The main goal of the work in this chapter is to investigate the effects of causal and non-causal explanations for AV actions in different driving scenarios with the intelligibility, accountability, and trust objectives in mind.

Intelligibility

H1.1: Intelligibility across explanation types Contrastive explanations are preferred by humans because humans generally expect a contrastive response when they ask questions (Miller, 2019; Mittelstadt et al., 2019). Therefore, we hypothesise that: *Why Not explanations would generally yield the best understanding of AV actions compared to Why, What If, and What explanations.*

H1.2: Intelligibility across scenarios As normative scenarios are more common in the real world, we assume that driving actions in these scenarios would be easier to comprehend compared to rarer scenarios, such as near-misses, emergencies, and collisions. Therefore, *we hypothesise that explanations would yield the highest level of understanding of AV actions in the normative scenarios.*

Accountability

H2: We contextualise accountability as the ability to recognise road/traffic rules violations and identify the road participants responsible for the violations. As *Why Not* explanations are useful for comparing outcomes, *we hypothesise that Why Not explanations will yield the best performance in accountability tasks.*

Trust

H3: We assume that intelligible explanations are useful in building a correct mental model of a system, and in turn, improve understanding and the perception of trust in the system. Hence, *we hypothesise that perception of trust and 'goodness of explanation' would correlate with the level of understanding of AV actions.*

3.2 User Study

In this section, we describe the user study conducted to investigate the effects of different explanations under different scenarios. We first describe the participants' demographics, and then the study design. The necessary approval to conduct the study was obtained from the University of Oxford's Research Ethics Committee.

3.2.1 Participants

We recruited 101 participants via the Prolific Academic platform ("Prolific", n.d.) and applied filters to include only individuals over age 18, living in the United Kingdom, and fluent in the English language. Participants did not have any language or reading disorders. To ensure quality results from the study, we ensured that the participants had completed at least 5 online surveys via Prolific before the time of the study, and received at least a 95% approval rate. There were 27 participants in the *Why* group, 24 in the *Why Not* group, 24 in the *What If* group, and 26 in the *What* group. 39 of the participants were males and 62 were females.

Their educational experiences ranged from high school diploma/A-level (29), enrolled for bachelor (12), bachelor's degree (48), to post-graduate degrees (12). 95 participants possessed at least one form of driving licence, while 6 did not. 49 participants indicated that they had prior experience driving on the left side, 14 of them had prior experience driving on the right, and 32 had prior experience driving on both sides. Asking participants how many days they drove in a typical week before the COVID-19 pandemic lock-down in March 2020, 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 for a week before the lock-down. Among those who drove, 26 of them had driven on minor roads, major roads, and motorways, while 40 of them had driven on only one of the three road classes.

We set Prolific to automatically time out participants who have spent more than 2 hours attending the online survey. This was stated in the instructions provided to the participants at the start. Dismissed participants were automatically replaced by Prolific. Participants took 38 minutes on average to complete the study. Each participant was paid £10 on completion.

3.2.2 Study Design

Before conducting this exploratory study, we examined the methodological aspects of related works (Binns et al., 2018; Lim et al., 2009), and adapted a combination of them. As 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 setup for participants to engage with certain driving scenarios involving an AV in order to learn the AV's behaviour, and subsequently get tested through a set of tasks. The learning process involved the presentation of different sequences of driving scenario images with explanations provided as captions. The testing process followed the same procedure as the learning procedure but the explanations were replaced by questions about the scenarios illustrated in the image sequences.

We investigated the effects of four types of explanations (*Why, Why Not, What If*, and *What*) through an online between-subject study with four groups. A between-subject design was chosen as against within-subject study because our study was very sensitive to carryover effects. We do not want participants to transfer driving and explanation experiences across runs.

Independent Variable

The independent variable was *explanation type* which involved four types of explanations; the *Why, Why Not, What If*, and *What* explanations.

Dependent Variables

We assessed four dependent variables, *task performance, perception of trust, traffic rule agreement*, and the *goodness of explanation*. *Task performance* scores were used as a measure of intelligibility and accountability. Trust questionnaires before (Pre-AV Experience) and after (Post-AV Experience) the exercise were used to capture participants' *perception of trust*. Other dependent variables, such as *traffic rule agreement* and the *goodness of explanation* were mainly used for triangulation purposes. Overall, the study was structured in three phases—Phase 1, Phase 2, and Phase 3 (see Figure 6.5).

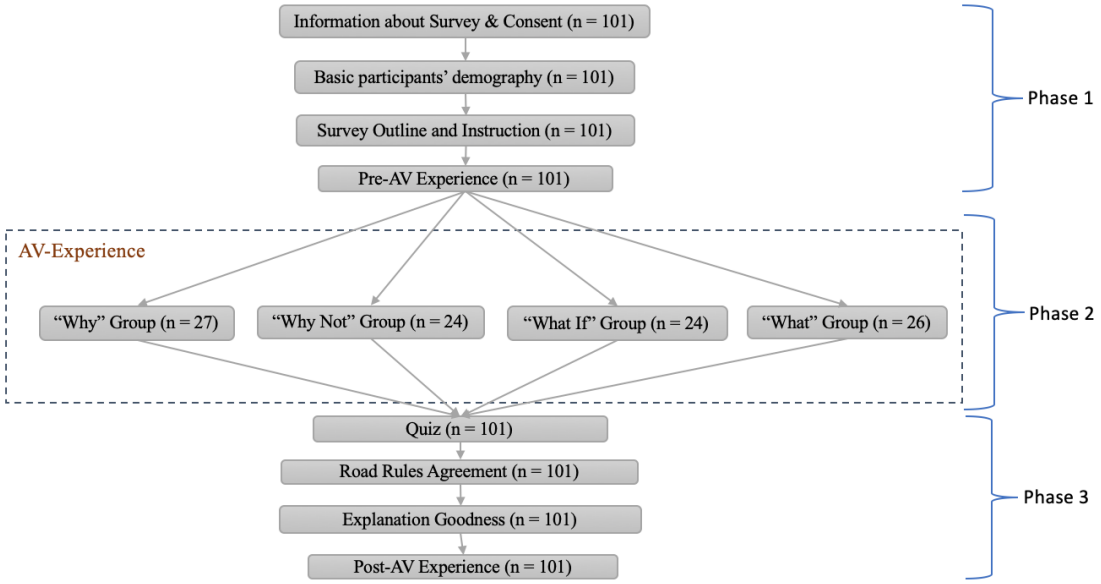


Figure 3.2: The user study has had three phases (1–3). The main part of the first phase is the pre-AV experience questionnaire on trust. The second phase is the actual experiment setup where participants were assigned to four different groups and provided with graphical illustrations of AV driving scenarios and textual explanations for each of the scenarios. The third phase includes all evaluations carried out, including task completion (in the form of a quiz), a road rules agreement questionnaire, an explanation goodness questionnaire, and a post-AV experience questionnaire on trust.

Phase 1

This phase comprised the first four sections as shown in Figure 6.5: Information and consent phase, participants' demography, instruction, and pre-AV experience questionnaire on trust.

In the pre-AV experience stage in Phase 1, participants were asked to respond to the first questionnaire on trust (the pre-AV experience questionnaire). This was presented to all groups to capture their perception of AVs, and especially, their perception of trust. The pre-AV experience questionnaire contained 8 questions with a 5-point Likert scale adapted from a psychometric trust scale recommended by Hoffman et al. (2018). The Hoffman trust scale was chosen after a review of other measurement scales in (Jian et al., 2000; Madsen & Gregor, 2000). The statements tested whether users agree that AVs are rule-abiding, predictable, reliable, safe, efficient, warying, effective, and adoptable by users. The statements were worded in the form: 'I currently have confidence in autonomous vehicles and I feel that they obey road rules and can respond appropriately to traffic situations.' Also, participants were asked to provide free responses about what they think of AVs: 'What do you think about autonomous vehicles? (e.g. trust, safety, reliability,...)'.

Phase 2

Participants were randomly assigned to four 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 still graphical images, illustrating driving scenarios, but with different types of textual explanations (i.e., *Why*, *Why Not*, *What If*, and *What* explanations) as captions, where each group consistently got one of the four types of explanations. Participants observed the driving scenarios by looking at the image sequences and reading the corresponding textual explanations (image captions) which explained the driving actions in the scenarios.

Driving Scenarios A scenario is represented with a sequence of images where each image is a frame at a time step. Image frames depict the actions of the different actors in a scene at each time instance. Based on the action categorisation in Ramanishka et al. (2018), the AV actions in each scenario could either be goal-oriented or stimulus-driven. 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

sudden stop or deviation decision due to the actions of other traffic participants, or obstacles on the vehicle's trajectory. These stop and deviate actions are categorised as stimulus-driven. We grouped scenarios as normative, near-miss, collision, and emergency. See Figure 3.3.

Figure 3.3: Driving scenario types and AV action categories. Concrete examples of AV actions demonstrated were left turn and lane merge. The blue vehicle is the AV while the yellow vehicle is an emergency vehicle. Figure 3a depicts a lane merge action (stimulus-driven) in a near-miss scenario. The AV had to slow down in order to let the other vehicle go first. Figure 3b is a lane merge action (goal-oriented) in a normative scenario where the red road sign indicated that the right lane has the right of way. Hence, the AV allowed the other vehicle to pass. Figure 3c is a left-turn action (goal-oriented) in an emergency scenario. The AV (though has the right of way) observed an emergency vehicle ahead and gave way right on time. Figure 3d is a left turn action (stimulus-driven) in a collision scenario. The AV proceeded after the red vehicle gave way, but the green vehicle from the side road suddenly crossed its path.

1. *Normative:* all road participants including the AV obey the road rules in this scenario type.

2. *Near-miss:* in this scenario type, a traffic participant violates the traffic or road rules, and the AV has to adjust to avoid a collision. Adjustments could be by steering, braking or a combination (Markkula et al., 2012).

3. *Collision:* In this scenario type, two or more vehicles (including the AV) crash into each other. This happens when one of the traffic participants suddenly violates road rules, and the AV fails to adjust accordingly to avoid an accident.

4. *Emergency:* This scenario type involves an emergency vehicle which could be 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 the two AV driving action classes (i.e., goal-oriented and stimulus-driven actions) in the different scenarios (i.e., normative, near-miss, collision, and emergency). Concrete examples of AV actions were left turns and lane merges (see Figure 3.3). There was an AV in every scenario, and it was always the only blue-coloured vehicle. The total number of scenarios designed in this phase was 24. Vehicles keep to the left in the United Kingdom, but we made vehicles keep to the right in the experiment, and we introduced new road signs in the scenarios in an attempt to place all participants on a seemingly levelled plane. Participants were asked to imagine that they were passengers in the AV, and that the explanations were generated by the AV. Meanwhile, the explanations were generated manually following the template in Table 3.1.

Explanation Generation To ensure consistency of explanation presentation forms within driving scenario classes, we created an explanation template (see Table 3.1) for the different scenario classes. The template was carefully designed to appropriately place the explanation elements for good intelligibility. The explanation elements included road rules, circumstances around the scenario in relation to other road participants, justification for actions, and outcomes.

Table 3.1: Schema for generating different types of explanations per scenario.

	Why	**Why Not**	**What if**	**What**
Normative	We [*x,y,...*] because [*describe circumstance*] and [*reference to one road rule/road sign*].	We didn't/can't/couldn't [*x,y,...*] because [*make reference to rule/sign*] when [*describe circumstance*].	If we [*x,y,...*] we will [*outcome of the decision*].	We are [*state action without reasons*]
Near-miss	We [*x,y,...*] because [*state the unexpected circumstance*] and [*state the offence with respect to a road rule/sign*].	We didn't/can't/couldn't/aren't [*x,y,...*] because [*state the unexpected circumstance*]. [*emphatically state the road rule or road sign meaning*].	If we [*x,y,...*] we will [*outcome of the decision*].	We are [*state action without reasons*]
Collision	We [*x,y,...*] because [*state the unexpected circumstance*], and [*state the offence with respect to a road rule/sign*].	We didn't/can't/couldn't/aren't [*negation of the action that led to collision*] because [*state why you are right*], [*state the unexpected circumstance*], [*reason for not adjusting immediately*]. [*emphatically state the road rule or road sign meaning*].	If we [*x,y,...*] we will [*outcome of the decision*].	We are [*state action without reasons*]
Emergency	We [*x,y,...*] because [*state the circumstance*] and [*state justification*].	We didn't/can't/couldn't [*x,y,...*] because [*state the circumstance*] and [*state justification*].	If we [*x,y,...*] we will [*outcome of the decision*].	We are [*state action without reasons*]

Phase 3

Phase 3 was an evaluation phase, set up to capture the effects of the explanations provided in Phase 2. Participants from all groups were asked to perform a set of tasks in the form of a quiz to assess their understanding of the explanations provided. They were also provided with different questionnaires, such as the road rules agreement questionnaire, explanation goodness questionnaire, and trust questionnaire. We explain these measurements in greater detail in the following section.

3.2.3 Measurements

We developed both objective measures (through a quiz) and subjective measures (through questionnaires).

Task performance as a measure for intelligibility and accountability

We assessed intelligibility of explanations and accountability using the participants' performance in a set of given tasks. This was under the assumption that highly intelligible explanations would yield an enhanced understanding of the AV behaviour, and in turn, result in a good performance in the provided tasks. Hence, the

participants were asked to perform some tasks in form of a quiz after interacting with the scenarios and explanations in Phase 2. It comprised 30 multiple-choice questions. Each question required the selection of one choice out of four choices where only one choice was correct. It also included scenarios that exhibited the different AV driving action classes under the different scenarios. 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. *Accountability*—the participant is presented scenarios without explanations and then asked to identify the road participants who violated or did not violate road rules.

2. *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.

3. *Situation Assessment*—a graphic about a traffic scenario is presented along with four statements about the presented scenario. Participants were asked to select one out of the four statements that mostly supported the actions and/or context in the scenario.

Trust questionnaires as a measure for perception of trust

In order to re-calibrate participants' trust in AVs, we asked the participants to respond to a questionnaire similar to the pre-AV experience trust questionnaire in Phase 1. However, all the questions in the post-AV experience trust questionnaire were conditioned on the explanations provided in the AV experience stage in Phase 2. The statements were designed to test whether the participants' views on AVs being rule-abiding, predictable, reliable, safe, efficient, warying, effective, and adoptable have changed after observing the driving scenarios. The statements were worded in the form: *'Based on the explanations provided by the autonomous vehicle in this survey, I have increased confidence in autonomous vehicles and I feel that they obey road rules and can respond appropriately to traffic stimulus'.* We framed all the statements in the questionnaire as follows:

1. *Rule abiding:* 'Based on the explanations provided by the autonomous vehicle in this survey, I have increased confidence in autonomous vehicles and I feel that they obey road rules and can respond appropriately to traffic stimulus.'

2. *Predictability:* 'Based on the explanations provided by the autonomous vehicle in this survey, I feel that the decisions of autonomous vehicles are predictable.'

3. *Reliability:* 'Based on the explanations provided by the autonomous vehicle in this survey, I feel that autonomous vehicles are very reliable. I can count on them to be correct all the time.'

4. *Safety:* 'Based on the explanations provided by the autonomous vehicle in this survey, I feel safe that when I rely on autonomous vehicles, I will safely get to my desired destination.'

5. *Efficiency:* 'Based on the explanations provided by the autonomous vehicle in this survey, I feel that autonomous vehicles are efficient in that they respond very quickly to their environment.'

6. *Wariness:* 'Based on the explanations provided by the autonomous vehicle in this survey, I am cautious about autonomous vehicles.'

7. *Effectiveness:* 'Based on the explanations provided by the autonomous vehicle in this survey, I feel that autonomous vehicles can perform their task better than a novice human driver.'

8. *Adoption:* 'Based on the explanations provided by the autonomous vehicle in this survey, I would like to start using autonomous vehicles for travelling.'

Participants were also asked to provide free responses about what they generally think of AVs. The question was *'Based on the explanations provided by the autonomous vehicle in this survey, what do you think about autonomous vehicles? (e.g. trust, safety, reliability,...)'.*

3.2.4 Other Measurements

We designed two other measurements, such as traffic rule agreement and the goodness of explanation questionnaires as additional dimensions to verify our results.

Traffic rules agreement questionnaire

To triangulate the task performance results, we created another objective means to assess participants' understanding of the AV behaviour and road rules. This was done by stating the most important road rules that applied during the AV-experience stage (or learning stage) and asking participants to rate their agreement with the rules on a 5-point Likert scale. An example of a road rule was: 'Yellow vehicles always have the right of way and every other vehicle has to yield in all circumstances.' We assumed that participants with good performance in the quiz would strongly agree with all the statements as the stated road rules were those rightly applied in the AV-experience stage.

Goodness of explanation questionnaire

To obtain specific feedback on the explanations provided, the participants were provided with seven statements testing for the basic properties of a 'good' explanation as discussed in (Hoffman et al., 2018; Holzinger et al., 2020; Miller, 2019; Mittelstadt et al., 2019); 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 in (Hoffman et al., 2018; Holzinger et al., 2020) and was adapted to fit our use case. The statements we used are:

1. *Understandability:* 'The explanations help me understand how the AV behaves.'

2. *Satisfaction* 'The explanations of how the AV behaved satisfied my curiosity.'

3. *Details:* 'The explanations of how the AV behaves are sufficiently detailed.'

4. *Completeness:* 'The explanations of how the AV behaves are sufficiently complete.'

5. *Actionable:* 'The explanations are actionable, that is, they help me know how to better interact with the AV in the future.'

6. *Reliability:* 'The explanations help me know how reliable the AV is.'

7. *Trustworthiness:* 'The explanations help me know how trustworthy the AV is.'

Further, participants were asked to provide free responses about what they did not like about the explanations, what they liked, and what they expect of a good explanation.

3.3 Quantitative Results

In this section, we present quantitative results from the task performance, perception of trust, road rules agreement, and goodness of explanations analysis, including our hypothesis tests.

Being a between-subject study, we performed Levene's test to confirm the homogeneity of variance assumption. This was to ensure that the within group variances—with respect to education level, previous driving experience of participants, among others—are equal for all groups. Levene's test passed with p-values ($p > 0.05$) for participants' education level, possession of a driving licence, length of driving experience, and driving wheel position experience (e.g., left, right, or both). Thus, we assume homogeneity of variance for all the identified potential confounding factors. We also checked the interaction effects between all of the aforementioned factors and explanation types with respect to task performance scores. No significant interaction effect was found, with p-values ($p > 0.05$). Hence, we went ahead to test our hypotheses.

3.3.1 Task Performance: Intelligibility and Accountability

We assumed that participants' performance in the tasks reflects how helpful the explanations were in enhancing their understanding of the AV's actions. Participants' performance scores did not violate homogeneity and normality tests, so analysis of variance (ANOVA) was used to check for statistical differences. Tukey's Honest Significance Difference (HSD) posthoc test was used in all cases where statistical differences were estimated. Tukey's HSD was used in order to obtain more confident results, reducing the chance for Type 1 errors.

Hypothesis H1.1—Intelligibility across explanation types

Why Not explanations would generally yield the best understanding of AV actions compared to Why, What If, and What explanations.

Explanation type significantly affected the participants' understanding of the driving scenarios as reflected in the quiz performances (quiz $F(3, 97) = 8.011, p < 0.001$). 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. This result supports hypothesis **H1** as the *Why Not* explanations overall yielded the best performance. Hence, hypothesis **H1** was not rejected (see Figure 3.4).

Hypothesis H1.2—Intelligibility across scenarios

Explanations would yield the highest level of understanding of AV actions in the normative scenarios. We wanted to see if there were performance differences between the four driving scenario groups (i.e. normative, near-miss, collision, and emergency). Applying the same technique in Section 3.3.1, we discovered that performance was significantly affected by some of the scenarios as follows: Collision (collision $F(3, 97) = 4.66,\ p = .004$) and emergency (emergency $F(3, 97) = 13.1,\ p < .001$), with the exception of the normative and near-miss scenarios. Participants generally performed best in the normative scenarios. The means and standard deviations were

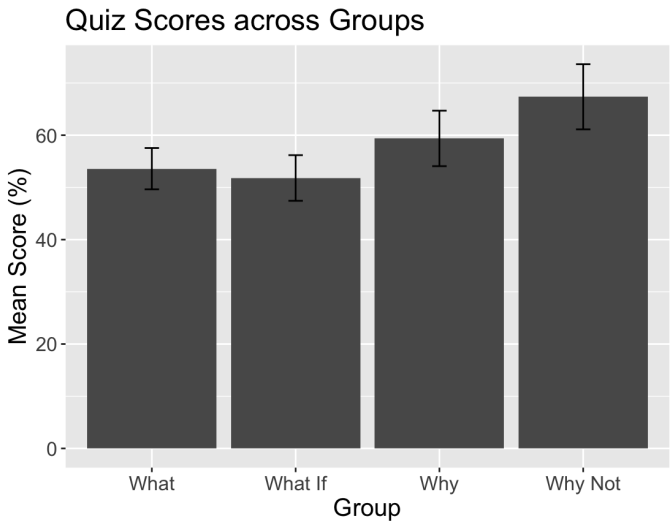


Figure 3.4: Performance of participants in the quiz. Participants in the *Why Not* group had the highest understanding of the AV's actions and *What if* group had the lowest understanding of the AV's action.

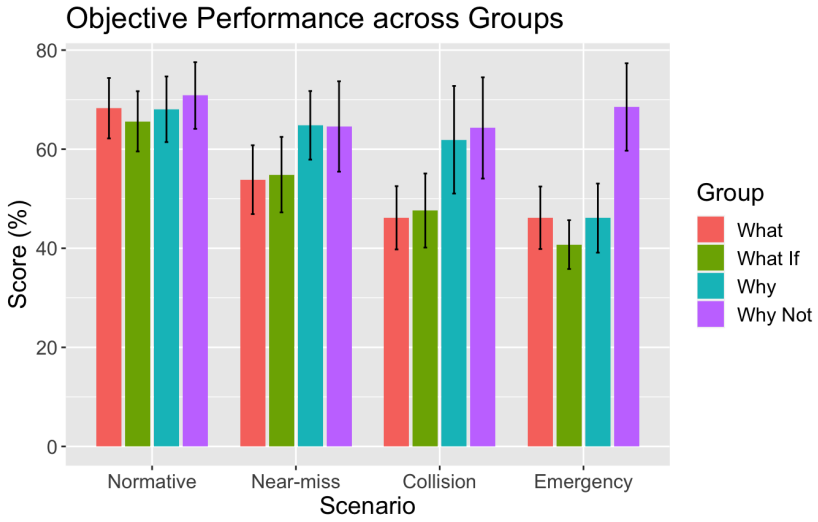


Figure 3.5: Quiz task performance in the different *driving scenario classes*. With the exception of the near-miss category, participants in the *Why Not* group consistently outperformed the participants in the other groups. Impacts of explanation type were greatest in the collision and emergency scenario.

($M = 5.46,\ 3.57,\ 3.85,\ 4.51,\ SD = 1.24,\ 1.14,\ 1.61,\ 1.74$) for normative, near-miss, collision, and emergency scenarios respectively (see Figure 3.5). Hypothesis **H1.2** has sufficient support, and hence, was not rejected.

Further, participants in the *Why Not* group performed best in the collision

scenarios, there was a significant difference between the *Why* and the *What* group (adjusted p-value = .04), *Why Not* and *What* group (adjusted p-value = 0.02), and *Why Not* and *What If* group (adjusted p-value = .04). Participants in the *Why Not* group performed best in the emergency scenarios, and there was a significant difference between *Why Not* and What groups (adjusted p-value < .001), *Why Not* and *What If* groups (adjusted p-value < .001), and *Why Not* and *Why* groups (adjusted p-value < .001).

Observing the score range (that is the maximum difference in scores in each scenario) across groups, the emergency and collision scenarios had the highest score range. Hence, explanations and explanation types are very critical in these scenarios. The order of scenarios in terms of explanation type importance is therefore: $Emergency > Collision > NearMiss > Normative$

Hypothesis H2—Accountability

Why Not explanations will yield the best performance in accountability tasks. We analysed scores based on the task categories (accountability, prediction, and situation assessment) to determine the category of tasks that participants performed better at. Analysing with ANOVA and Tukey's posthoc tests, we discovered significant differences across the groups in the accountability tasks (accountability $F(3, 97) = 6.96,\ p < .001$). The accountability task's mean score and standard deviation were 5.32 and 1.83, respectively. For the accountability tasks, participants in the *Why Not* group had the best performance having significant differences with *What* group (adjusted p-value = .01), *What If* group (adjusted p-value < .001), and *Why* group (adjusted p-value = .007). Hence, Hypothesis **H2** has sufficient support and was not rejected.

In addition, we analysed the data from the prediction and situation assessment tasks. There were significant differences across groups in the prediction tasks (prediction $F(3, 97) = 4.03,\ p = .01$) and situation assessment tasks (situation assessment $F(3, 97) = 5.62,\ p = .001$) with ($M = 6.23,\ 5.81,\ SD = 1.44,\ 1.98$) respectively. This implies that participants performed best in the prediction tasks.

Participants in the *Why Not* group had the best performance with significant difference between *What* group (adjusted p-value = .01), *What If* group (adjusted p-value = .04), and *Why* group (adjusted p-value = .04). Participants in the *Why* group had the best performance in the situation assessment tasks with a significant difference between the *Why* and *What If* group (adjusted p-value = .003), and *Why Not* and *What If* groups (adjusted p-value = .02).

3.3.2 Perception of Trust

Hypothesis H3—Perception of Trust

Perception of trust and 'goodness of explanation' would correlate with the level of understanding of AV actions.

We computed the difference in means for each participant's responses to the eight Likert statements in the pre-AV experience and the post-AV experience questionnaires. We checked whether the differences were significant across the groups. Our results indicate the absence of significant statistical differences and showed no correlation between task performance scores and trust difference ($rho = 0.037$, $p = .71$). There was no sufficient support for Hypothesis **H3**, hence, Hypothesis **H3** was rejected.

In addition, participants' perceptions of the trust factors in AVs mostly declined in the post-AV experience stage. The number of participants, who indicated that they would like to start using AVs for travelling, reduced in the post-AV experience stage, see Figures 3.6 and 3.7.

3.3.3 Other Quantitative Results

Goodness of Explanation

ANOVA test indicated a significant difference in explanation goodness rating across groups ($F(3, 97) = 10.0, p < .001$). Means and standard deviations of the goodness of explanation ratings were: (M = 3.83, 3.34, 3.25, 2.83, SD = 0.35, 0.65, 0.77, 0.88) for *Why, Why Not, What If,* and *What* groups respectively. The highest mean rating was from the participants in the *Why* group.

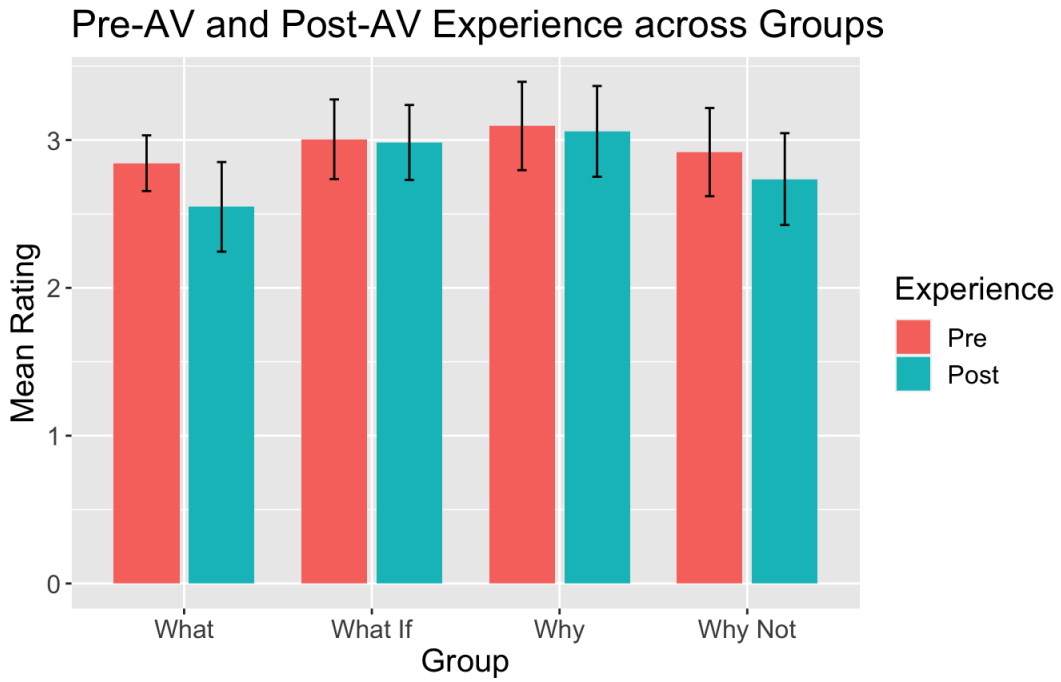


Figure 3.6: Trust factors evaluated through the pre-AV (pink) and the post-AV experience questionnaire (green) with 5-point Likert scales. The explanations without causal attributions (i.e.,*what group*) had the highest decline.

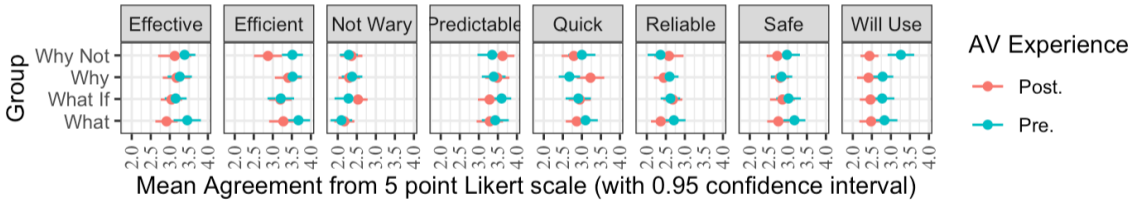


Figure 3.7: Mean values for each of the trust factors represented in the pre-AV experience (represented with the green plot) and post-AV experience (represented with the pink plot) questionnaires. Each box caption represents a trust factor. The pre-AV experience factor had higher values but there was no significant difference between the pre-AV and post-AV experience factors across the groups.s

Tukey's test revealed a significant difference between the *Why* and *What* group (adjusted p-value $< .001$), *Why Not* and *What* groups (adjusted p-value $= .042$), *Why* and *What If* groups (adjusted p-value $= .015$), and *Why Not* and *Why* groups (adjusted p-value $= .048$). No correlation was observed between the explanation goodness ratings and the quiz scores (rho $= 0.19$, p $= .051$). See Figure 3.8 and 3.9.

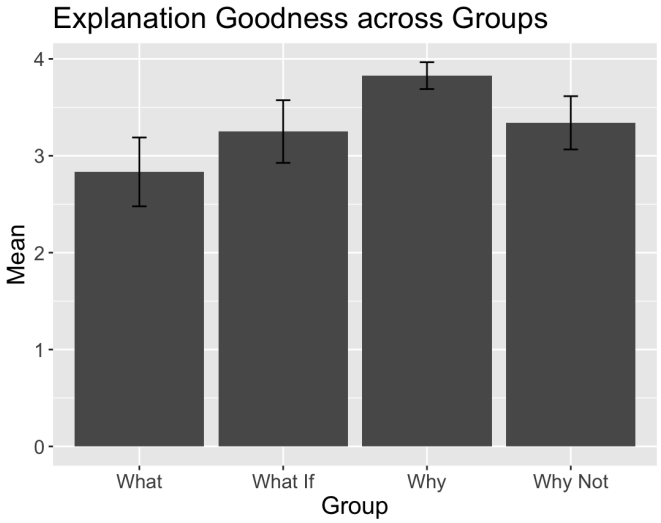


Figure 3.8: The figure describes the mean goodness of explanation values (i.e. mean Likert scale values) for all of the goodness factors in goodness of explanation questionnaire. The y-axis indicates the mean goodness values while the x-axis indicates the respective groups. Participants in the *Why* group gave the highest explanation goodness rating.

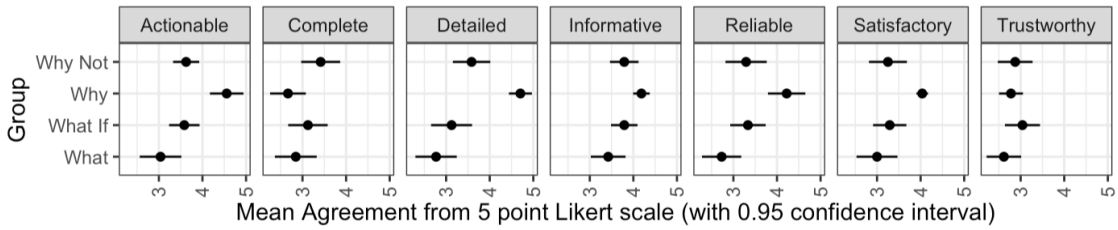


Figure 3.9: The figure describes the mean goodness of explanation values (i.e. mean Likert scale values) for all of the goodness factors in goodness of explanation questionnaire. The x-axis indicates the mean explanation goodness values while the y-axis indicates the respective groups. The captions indicate the different explanation goodness factors from the questionnaire. Participants in the *Why* group gave the highest rating.

Driving Rules Agreement

Using Pearson's correlation coefficient, we checked the correlation between the mean road-rules agreement ratings from the participants and their quiz scores. The result showed that there was a weak positive correlation between the two variables ($rho = 0.34,\ p < .001$). 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* groups respectively (see Figure 3.10 and 3.11). This indicates that the *Why Not* and *What If* group members understood the road

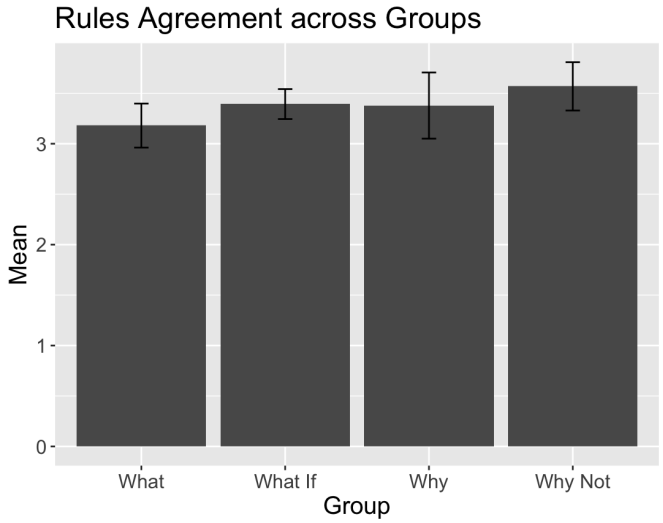


Figure 3.10: Rule agreement performance from the rule agreement questionnaire. Participants in the *Why Not* group best understood the rules as they mostly agreed with the road rules.

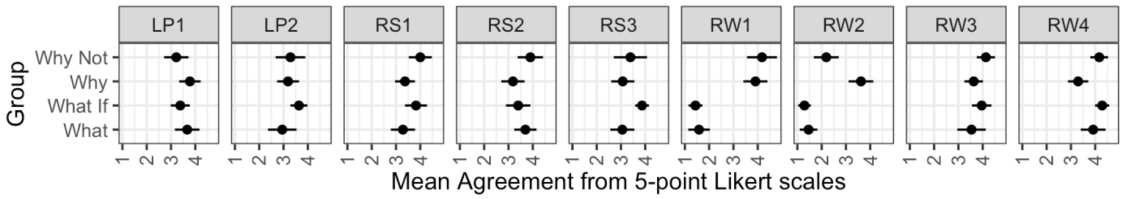


Figure 3.11: The figure describes the mean agreement values (i.e., mean Likert scale values) for each of the stated rules in the rule agreement questionnaire. The x-axis indicates the mean agreement values while the y-axis indicates the respective groups. The captions indicate the type of road rule stated. LP: Lane Position rules, RW: Right of Way rules, RS: Road Signs

rules better in the AV-Experience.

3.4 Qualitative Results: Themes and Reflections

In addition to the 5-point Likert scale questionnaires, the participants were asked to provide free responses in the pre-AV experience, and the post-AV experience trust questionnaires, and the explanation goodness questionnaire.

3.4.1 Perception of Trust: Pre-AV Experience and Post-AV Experience

The free response questions asked as follow-up questions to the pre-AV experience and post-AV experience questionnaire are as follows:

Pre-AV Experience Questionnaire: 'What do you think about autonomous vehicles? (e.g. trust, safety, reliability,...)'

Post-AV Experience Questionnaire: 'Based on the explanations provided by the AV, what's your thought on autonomous vehicles (e.g. trust, safety, reliability,...)'

We performed a thematic inductive analysis on the qualitative data captured to assess trust. Excerpts were taken from participants' comments in the pre-AV experience and the post-AV experience in order to better explain the derived themes. Generally, there was an indication of a decline in the participants' perception of trust. We discuss the themes under two broad categories: distrust supporting themes and trust supporting themes. See Figure 3.12 for the frequency plot of the trust and distrust comments.

Distrust

Participants' perceptions of trust did not improve. In fact, a decline was noticed instead. See distrust supporting comments below.

Unwillingness to give-Up control: Though some participants seemed to have an increased understanding of the AV after the AV experience stage, their perception of trust did not improve. They still preferred to have full driving control over the vehicle. See an example comment:

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

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

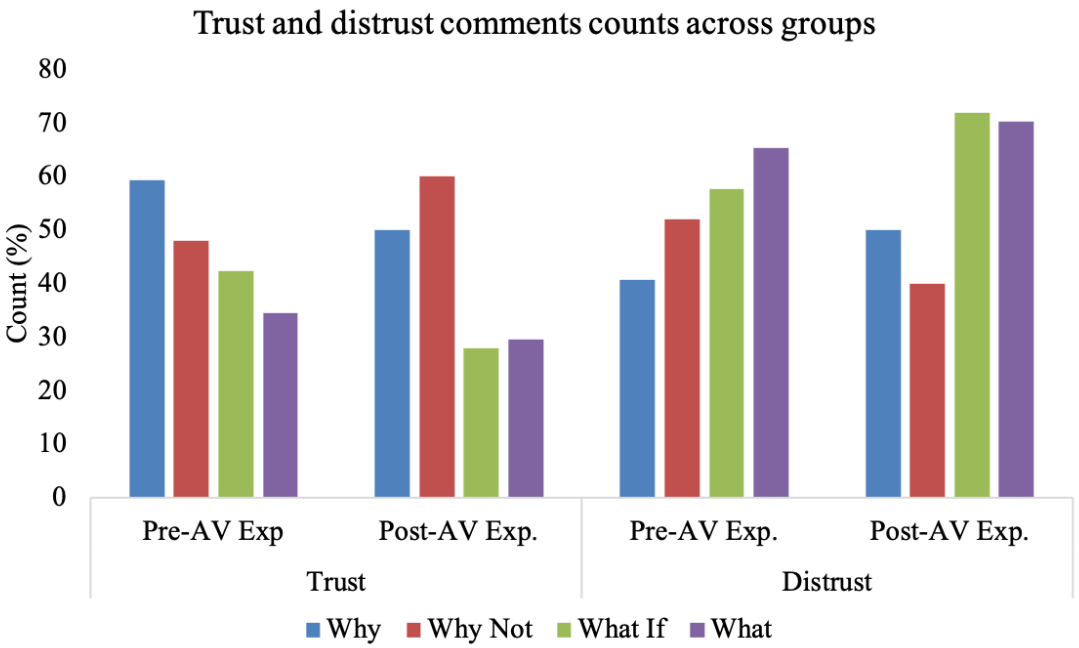


Figure 3.12: Frequency of negative (distrust) and positive (trust) comments about trust in AVs. The y-axis indicates the frequency in percentage, while the x-axis indicates the pre-AV and post-AV experiences along with trust and distrust comments. Only the *Why Not* group had increased positive comments in the post-AV experience questionnaire.

Too early to be trusted: The nascent nature of AV technologies made some participants wary about their current capabilities, thinking that they are still too early to be trusted:

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

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

Worry about reliability and robustness: Though many participants agreed that the AV complied with road rules in the post-AV experience, there were worries about the AV's reliability and its response to unpredictable situations. Participants were not sure of the efficiency of the take-over process in such unpredictable situations:

> 'I have four big concerns: 1) How reliable is the current technology; Could they spot and avoid a child dressed in dark clothes who runs out

> into the road during a thunderstorm as effectively as a human? [...] 3) Reliability. What if the vehicle malfunctions; A big malfunction is easy to spot, but what about something subtle? There might be no obvious problem until there's an accident [...]'–NG (pre-AV experience).

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

Track records of accidents: Apparently, previous records of accidents in the news, and the few collision examples (due to traffic offences by the other road participants) shown to the participants in the AV-Experience stage of the study might have increased participants' doubts:

> 'I would not use one until they are firmly established into society and have a proven track record of safety.'–DA (post-AV experience).

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

Worry about environment, security, and prioritisation: Participants thought that the dynamic nature of driving environments, prioritisation of road participants (not just emergency vehicles as shown in the scenarios) and security of the AVs were crucial. This might have affected their trust factor ratings in the study:

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

> 'It needs more consideration about different traffic circumstance including priority feature.'–GA (post-AV experience).

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

Ethical concerns: Some participants raised ethical concerns about the design process of AVs:

> '[...] Ethics. Who programs the vehicle to make choices[...]'–NG (post-AV experience).

> '[...]It has also not gone into any depth about moral decisions, [...]'–BJ (post-AV experience).

Trust

Some of the participants in the groups that received causal explanations indicated that the explanations were helpful in explaining the reasons behind the driving decisions of the AV and therefore somewhat trusted the AV.

Explanations enhanced trust: A participant in the *Why Not* group had a change of view from distrust to trust.

> 'I don't understand how they can react quickly enough in an emergency situation to be safe. I am very wary!'–KP (pre-AV experience).

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

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

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

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

> 'I have more trust in their decision making - the main worry is the human drivers! If all cars were AVs then I would have a lot more faith as this takes a lot of unpredictability out of the equation.'–LG (post-AV experience).

Improved productivity and appeal for publicity: Some also thought AVs could be well-suited for long journeys, and that designers and manufacturers needed to promote this benefit better for widespread adoption:

> 'I think they would be an excellent way for me to make more out of my day. 2 hours commuting time where I may be able to do other things than concentrate on the road'–CA (pre-AV experience).

This participant remained positive and stresses the need for more AVs.

> 'I think that they can be trusted, but a lot more data needs to be collected before we get there. Also I think it will be safer for all cars to be autonomous rather than some [autonomous] and some normal drivers'–CA (post-AV experience).

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

> 'I think AV manufacturers and companies should promote [AVs] more to the public to get better widespread perceptions'–FR (post-AV experience).

3.4.2 Other Qualitative Results: Goodness of Explanations

The free response questions asked as a follow-up to the 5-point Likert scale are as follows:

- '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 grouped the responses into themes across groups based on reoccurring comments. Participants generally prefer short explanations with sufficient information. Comments were made on the explanation presentation style. For example, some participants suggested that explanations be provided as bullet points and that the AV should provide more details on road signs when explaining scenarios. A summary of the themes is presented in Table 3.2.

Table 3.2: Common themes from participants' comments about the explanations provided to them.

	Limitation	Strength	Suggestion
Why	- Not convincing enough to trust AVs in unexpected situations So many outcomes - Information overload - Ineffective communication of speed priority	- Clarity - 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 - Reaction in conflict
Why Not	- Information overload - Situational report and not mechanistic - Road signs unexplained - Not enough clarity	- Simple and easy to follow - Short and snappy - Highlighted and detailed - 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

3.5 Discussion

The results from the study indicate that Hypothesis **H1.1** holds as participants in the *Why Not* group generally performed better than participants in the *Why, What If*, and *What* groups. This supports our claim that humans generally expect a contrastive response when they demand an explanation for an outcome that differs from their decisions (Miller, 2019). Generally observing from Figure 3.8 3.9, 3.6, and 3.12, explanations with causal attributions (i.e., Why, Why Not, What If explanations) are more preferred to those without causal attributions (e.g., What explanations). Observing the score ranges across groups, the emergency and collision scenarios had the largest score ranges. Hence, explanations and explanation types are very critical in these scenarios, especially for incident investigation purposes (Figure 3.5).

Hypothesis **H1.2** was supported by the results as the participants overall had a higher level of AV action understanding in the normative scenario through explanations. Participants in the *Why Not* group performed better than the other groups in the emergency scenarios. However, this was not the case in the near-miss scenarios.

Participants in the *Why Not* group produced the best performance in the accountability tasks in support of Hypothesis **H2**. In addition, participants in the *Why Not* group agreed the most with the road rules stated after the quiz. This implies that the *Why Not* explanations were successful at making the participants understand the road rules that governed the AV in the scenarios. While the rule agreement means correlated with the quiz scores, no correlation was observed between the performance scores and the pre-AV experience and post-AV experience difference (i.e. trust metric). Hence, there is no sufficient support for Hypothesis **H3**. We further discuss some interesting findings from the study.

3.5.1 Intelligibility

Explanations have been proven in (Hagras, 2018; Hayes & Shah, 2017; Selkowitz et al., 2017) to increase intelligibility in autonomous systems. This was made evident in our experiment where participants who received explanations with causal attributions performed better in the different task categories. In particular, the contrastive explanations (i.e. response to a *Why Not* investigatory query) were identified to provide the highest level of intelligibility in the AV as the explanations led to the best performances in most of the tasks presented to the participants. Accurate accountability (i.e., correct assignment of responsibilities), situation assessment, and action prediction are an exhibition of a possible display of a user's correct mental model of the driving situations and the AV's behaviour (Hoffman et al., 2018). In general, the outcome of the study corroborates Glymour (1998) claims that explanations with causal attributions are more effective in explaining AI systems (Glymour, 1998), and in this case, AVs' behaviours. Hence, designing causal explanations with the participants' feedback (such as being concise, provision of sufficient reason, and more interactivity) in mind will be very helpful for the development of future AVs.

3.5.2 Accountability

Intelligibility of explanations may lead to better accountability. The accountability tasks presented to the participants in the quiz section required that the participants assess situations and identify road participants who did not comply with the road rules in each of the scenarios. As mentioned in Section 3.3.1, the *Why Not* explanations were the most effective in the accountability quiz tasks. We conclude that explanations with causal attributions (and in particular contrastive ones) are helpful in improving accountability in autonomous driving.

3.5.3 Perception of Trust

While explanations might have some relationship with trust (Pieters, 2011; Pu & Chen, 2006), they may not directly or instantaneously create or improve trust as observed in this study. In the goodness of explanation evaluation, the *Why* and *Why Not* groups had the highest mean ratings for most of the explanation goodness factors. This was also reflected in their performances in the objective (quiz and road rules agreement) tasks with *Why Not* group having the highest performance and then the *Why* group. However, participants in the *Why Not* group did not have a significant positive change (where they exist) in each of the trust factors they rated before and after the experiment (i.e. the pre-AV experience and post-AV experience stages). For example, in the *Why Not* group, the mean rating for all the trust factors declined in the post-AV experience except for predictability and reliability. From the qualitative data, many participants who expressed distrust in the pre-AV experience stage also indicated distrust in the post-AV experience stage but with reasons that relate to their enhanced knowledge and experience of the AV in the post-AV experience stage. See example responses:

> 'They sound like a good idea as it avoids human error, however, I worry about what would happen if there was a technical malfunction and the driver is not in control of the vehicle. I don't know enough about them to be able to make an informed decision regarding reliability, safety.'–GH (pre-AV experience)

> 'The explanations indicate that the AV always follows the rules of the road, however, I'm concerned regarding the outcome when other road users don't follow the rules.'–GH (post-AV experience).

Furthermore, despite having the best quiz performance in the *Why Not* group, trust factors ratings were better in the *Why* group. However, there were more trust comments from the *Why Not* group in the post-AV experience. Overall, there was no significant positive change in the perception of trust but rather a decline.

3.5.4 Regulations and Standards

Regulators have a role to play in ensuring trustworthy autonomous vehicles. Some of the concerns raised in the study touched on effective testing by relevant stakeholders to ensure safety. As an example, a participant commented on the need for long-term testing to assure safety:

> 'I think they can be reliable if tested for a long period of time but at the moment I don't trust them enough'–PO (post-AV experience).

Trust might improve when the public is provided with the assurance of safety and reliability not only from the manufacturers but also regulators.

3.5.5 Prior Experiences

Lastly, prior experiences of similar systems to AVs—even when the systems' operational modes differ—may influence users' perception of the new system cf. Wason's selection task (Cox & Griggs, 1982). For example, in (Lim et al., 2009), participants with prior experiences of similar tasks performed poorly in the given task.

As some of the participants in our study had prior experience driving both on the left and right side traffic on minor, major and motor-ways roads, we expected a positive trend (e.g., better task performances) from this category of participants. However, there was no significant positive trend identified. This might be as a result of the new road signs we introduced in some of the scenarios, and/or the right-hand drive that was enforced instead of the UK's left-hand driving to put

all participants on a levelled plane. We suggest that prior driving experiences and educational experiences might not have a significant effect on users' understanding of a similar system with different dynamics, especially when the users are fluent in the language of communication and do not have a learning disability.

3.6 Conclusion

We have described a study that compared the provision of explanations with causal attributions (i.e., *Why, Why Not*, and *What If*) and explanations that do not provide causal explanations (e.g., *What*) explanations. We designed an online study that employed these explanations to explain autonomous driving scenarios illustrated in sequences of graphical images. We asked participants to perform specific tasks. The performances from the tasks were used as a measure of the participants' understanding of the explanations provided for the AV's actions in the scenarios (intelligibility). Our findings disclose that providing causal attributions, and in particular, contrastive (or *Why Not*) explanations, can improve users' understanding of AVs, enhance accountability, and provide better interactions with AVs. This, however, might not necessarily improve the users' perception of trust in AVs.

There are a few limitations in the work that is worth highlighting. First, participant selection was restricted to those who live in the UK and are fluent in the English language. Driving rules differ between countries and it could have been good to have diverse participants with different driving experiences, and different levels of English language proficiency. Also, the number of female participants was almost twice the number of male participants. Generally, reports show that there are more male drivers (Abby, 2020).

Second, the study involved the use of hypothetical scenarios only. Participants were asked to imagine they were inside an AV in a graphical illustration in an attempt to create a realistic feeling. It, therefore, lacked the first-person consequences and the significance of a real-world decision. The explanations were generated manually following a defined template. While they looked similar to what an autonomous system might generate, they may not be a direct substitute.

Finally, explanations were only communicated in textual form and did not allow for personalisation and conversation. Ideally, users should be able to choose in what form they want explanations to be presented to them and ask follow-up questions if necessary. This is however outside the scope of this thesis.

Generally, this work is an important first step towards understanding the impact of explanations in AVs. In the next chapter, we will discuss the different representations and data structures that would facilitate the development of explanation algorithms in autonomous driving.

4

Explanation Generation: Representation and Algorithms

Contents

4.1 Introduction

As autonomous vehicles gain increasing attention and are getting more sophisticated, it is imperative that humans are able to interact with them effectively to fulfil tasks (Langley et al., 2017), while appropriately calibrating their trust in the process. In response, there have been recent advances around explainable agencies. Agents

are being equipped with the capability to explain their behaviours/decisions to humans. While these developments are plausible, architectural decisions, often ignored, are critical for obtaining more fruitful outputs.

The architectural design choice for autonomous vehicles has an important role to play in the derivation of explanations in autonomous driving. Two general options from which design choices are made include end-to-end and modular pipeline architecture (Tampuu et al., 2020). In the end-to-end pipeline, input streams are collected from the perception system (e.g., video frames) and control actions (e.g., trajectory and speed) are predicted as outputs. In other words, navigation decisions are only based on perception inputs passed to a black-box model which predicts actions. Although this approach is gaining increasing attention in the autonomous driving community due to the prevalence of high-performing vision-based deep models, it is disadvantaged by its tightly coupled nature, high complexity, and lack of transparency. This could, in turn, complicate system audit. In the modular pipeline, there is a better separation of concerns in that driving tasks are performed in stages by individual autonomous vehicle components such as those described in Chapter 2. The core components include the perception system, localisation system, planning system, and vehicle control system. Although not conventionally categorised as a core process in the driving task, the system management component is very critical as it sits across all the other components and logs errors and data for warnings and user engagement purposes. Moreover, this component is key for explainability as it sits between other core operations of the AV and the human. The research question we seek to answer is: **How can intelligible posthoc explanations be generated automatically for AV actions? Specifically, how do we obtain explainable representations, data structures, and algorithms for the generation of these explanations?**

In this chapter, we will discuss in detail the key considerations for an explainable AV in general (Section 4.2). We will also propose a conceptual explainable AV architecture (Section4.3) and representations (Section 4.4) for achieving explainable AVs. Using two case studies: collision risk explanations, and driving actions

explanations, we would build on the proposed representations to design algorithms for intelligible explanation generation (Section 4.5).

4.2 Fundamental Considerations for Explainable AVs

To effortlessly achieve explainability, we desire AV architectures to be:

- Non-complex: its inner workings should be expressible in clear natural language.
- Transparent: the decision-making process of the AV and its explainer should be traceable.
- Facilitate easy auditing of the AV: ultimately, AV auditors should benefit from an easier auditing process brought about by the transparency in the AV's functioning and the AV's governance operations.
- Facilitate easy incident investigation: data logs and explanations logs should be available in understandable forms for posthoc incident investigations.

Based on the aforementioned expectations, we propose a conceptual framework for explainable AVs which build on the modular pipeline. This design can serve as an initial template for the design of future explainable AVs.

4.3 Explainable AV Conceptual Framework

We suggest a conceptual framework of how we envision the core AV operations and components fitting together for intelligible explanation provisions purposes that would benefit different stakeholders (e.g., incident investigators and passengers). A diagrammatic illustration of this framework is shown in Figure 4.1. Based on the figure, the perception system provides a digital 3D representation of the world/environment, including information about object detection, detection and tracking uncertainties, and object location, and passes them to the behaviour

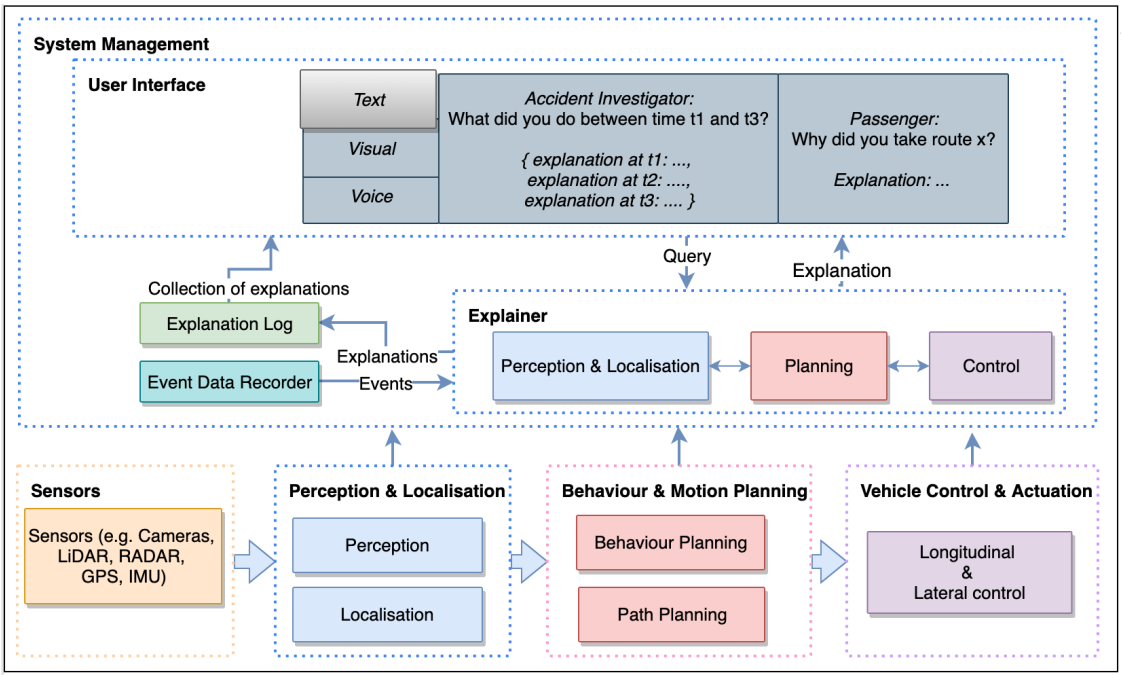


Figure 4.1: A conceptual framework for explainable AV following the perception, planning, and control paradigm. The explainer receives data from the perception and localisation system, planning system, and vehicle control system. This data can either be stored with time stamps and provided all at once as an explanation log to accident investigators, or each explanation is provided as it is generated. The framework provides a multi-modal interface for interacting with the explainer.

and motion planning system, which makes decisions by estimating risks, finding mitigation actions, and producing a trajectory. The perception and localisation operations are fused together as both receive input directly from the sensors (e.g., LiDAR, radar, GPS, IMU). The output from the behaviour and motion planning system informs the vehicle control and actuation system on the continuous signal type to send to the actuators to alter the vehicle dynamics in order to achieve a driving effect e.g., reacting to an observation. Each of these operations feeds directly to the explainer and the event data recorder (EDR) in the system management compartment. The EDR organises its data and can provide its organised/structured data to the explainer. The explainer can generate explanations from a combination of data from the different operations and the EDR. Explanations can be requested in batches in the form of an explanation log by specifying a time range or can be requested in a conversational manner (i.e., one query at a time). The explanation log can provide a process-based explanation useful for incident investigation and

effective AV governance. Process-based explainability in the AV context is concerned with providing information that facilitates the independent assessment of the entire operations and governance of the AV. Process-based explainability considers perception, decision, and action data, including the governance processes of the entire AV operation for explanations. This makes it possible to reconstruct an event or accident immediately after it happens, significantly reducing the time to provide recommendations for future improvements. Explanations may also be provided specifically with the purpose of understanding and rationalising the actions/decisions of an AV. The ICO refer to these explanations as outcome-based. While the proposed conceptual framework can afford process-based explainability for stakeholders like incident investigators, the rest of this work focuses on outcome-based explainability to benefit passengers.

We go a step further to explore how the AV can represent its data in order to provide intelligible explanations following the envisioned conceptual modular explainable AV framework. One way to go about this is to leverage trees and graphs, which are inherently interpretable and expressive.

4.4 Tree-based Representation

We translate the outputs from the different AV operations in Section 4.3 into high-level semantics that inform a new approach to the explanation generation process. We propose a new transparent approach to explanation generation that utilises a tree structure. This approach builds on the risk object identification technique in driving scene (C. Li et al., 2020) and traffic objects representation using scene graphs (L. Kunze et al., 2018). After a careful analysis of different driving scenarios, we identified three variables required to provide an intelligible explanation. This includes a set of road rules (R), observations (O), and actions (A). The processed output from perceptions is referred to as *observations* and can be represented with a scene graph. The resulting effect of control signals on the AV is referred to as *actions* and occurs on a temporal scale. The standard traffic constraints which are used during behaviour/motion planning are referred to as road rules (see Figure 4.2).

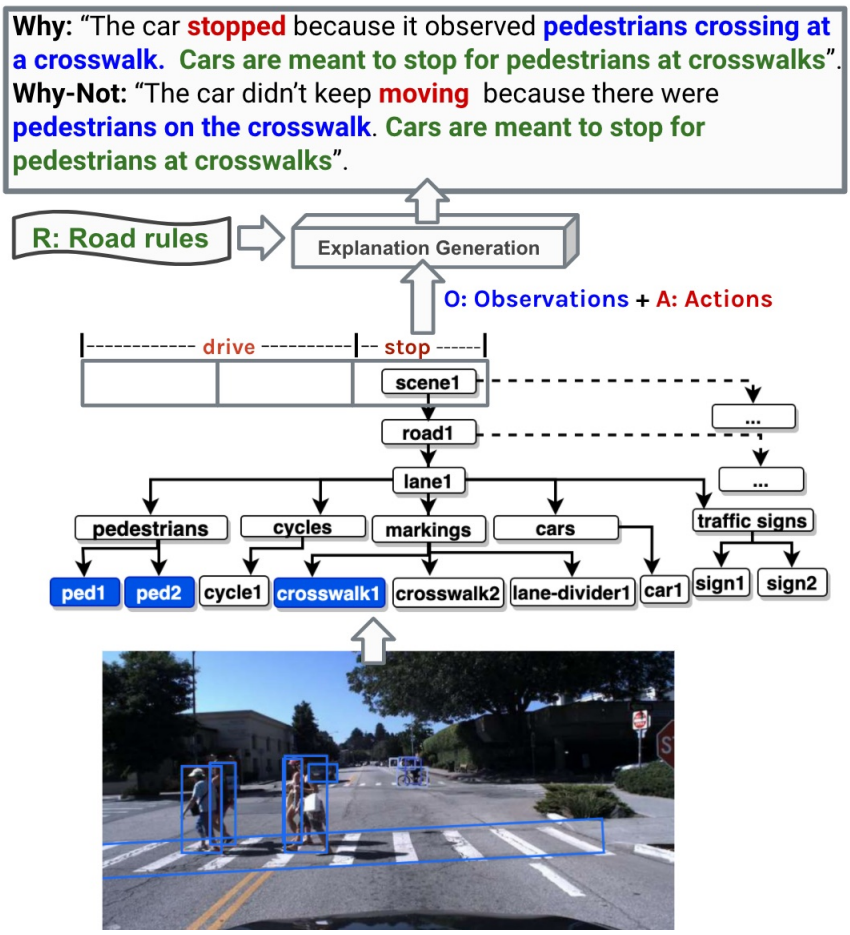


Figure 4.2: Interpretable Representation. Different types of explanations (e.g. *Why* and *Why Not*) are generated from the underlying representations of actions (A), observations (O), and road rules (R). The observations and actions are obtained from scene graphs representing various frames from a driving 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.

The observations recorded in the scene graphs, the AV's actions, and the road rules are combined to generate explanations. In this thesis, we only focus on how this combination can generate different explanations. Thus, we propose an interpretable tree-based representation for generating different types of explanations based on *observations*, *actions*, and *road rules*. See Figure 4.3 for how an explanation is generated for an example scene.

Explanation E is represented with a variable size tuple ($E = \langle ... \rangle$). For example,

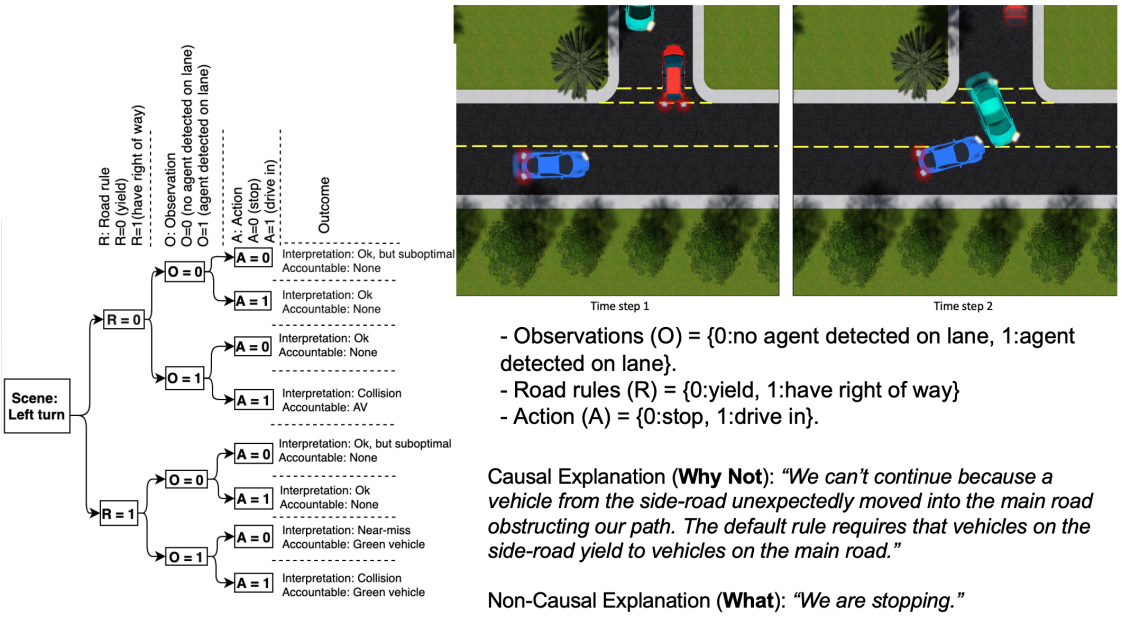


Figure 4.3: The tree on the left is the underlying tree-based representation for explanation generation for the scene on the right. The tree is constructed with key variables: road rules (R), observations (O), and actions (A). Different types of explanations are generated through different traversals of the tree. We can assess accountability through the explanations, or by interpreting outcomes for each path in the tree.

$E = \langle A, O, R \rangle$ will yield an explanation of the form:

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

In each driving scenario, a tree can be created from the observations in the scene graph, actions, and road rules (see Figure 4.2). The tree is traversed to collect the values for each member of the tuple needed to form E. Figure 4.3 (left) shows an example of a constructed tree using R, O and A for the left turn scenario (right).

Note that in this representation for explanation generation, we have used high-level scene information (e.g., agent type and agent action), which we assume that the scene understanding model (perception and localisation system) would provide. This limits the utility of the generated explanations to enhancing AV end-users' understanding, which is useful for passengers—the key stakeholder of consideration in this thesis. For an incident investigation task, we must be able to go under the hood and obtain the logic of the perception or scene understanding system. This is beyond the scope of this thesis.

Using two case studies, we will explain how these tree structures can be used to generate explanations. We only focus on the generation of explanations with causal attribution in this chapter (i.e., *Why, Why Not,* and *What If*). As *Why Not* is a trivial case of *What If* in terms of implementation, we would only illustrate the algorithms for *Why* (factual) and *What If* (counterfactual) explanations. Each case study is laid out in subsections as follows: preliminary, problem statements, and algorithm design for factual and counterfactual explanations.

4.5 Case Study 1: Transparent Collision Risk Assessment

The purpose of this case study is to introduce a simple case of explanation generation from a tree-based representation. We formulate a collision risk prediction task. We fit different tree-based models for risk prediction and illustrate how natural language explanations can be generated from the models. Collision risk prediction is an important task in autonomous driving for planning. While this case study mainly focuses on illustrating how explanations could be generated from tree-based models, AV developers could use this explanation generation method to facilitate model debugging and enhancement processes when the model concerned is tree-based. We set out this section by first defining risk metrics, and then describing the algorithms to generate explanations for a collision risk prediction task.

4.5.1 Preliminaries

Risk Metrics

Time-to-Collision (TTC) at an instant t is defined as 'the time that remains until a collision between two vehicles would have occurred if the collision course and speed difference are maintained' (Mahmud et al., 2017). This definition of TTC implies that if the speed of the following vehicle is larger than that of the leading vehicle, a collision will occur and ignores any potential conflicts due to acceleration or deceleration changes. The Modified Time to Collision (MTTC) (Ozbay et al., 2008) takes these limitations into account and provides a better risk assessment

that takes cognisance of varying speeds and accelerations. However, just like TTC, MTTC only works for rear-end collisions and thus fails to model a real world environment in which collision can occur in any direction. Hence, Ward et al. (2015) extended TTC by estimating it on a 2D plane rather than a 1D manifold, and also introduced the concept of looming. This planar TTC is calculated by assuming constant acceleration in contrast to the constant speed assumption. In planar TTC, time to Collision values can be calculated in two ways which are: the first order TTC (T_1), which is the TTC value calculated when change of closure rate is omitted. T_1 assumes a constant closure rate between the vehicles. The second order TTC (T_2), accounts for changes in closure rate.

Consider d_{ij} to be the distance between the closest points of an AV (a_{v_i}) and the agent (a_j) and $\dot{d}_{ij}$ and $\ddot{d}_{ij}$ to be its first and second derivatives respectively. If

$$T_1 = \frac{-d_{ij}}{\dot{d}_{ij}} \tag{4.1}$$

is the first order TTC where the closure rate is omitted, and

$$\Delta = \dot{d}_{ij}^2 - 2\ddot{d_{ij}}d_{ij} \tag{4.2}$$

is the discriminant of the second-order case, then the formula to calculate the planar TTC (T_2) is as follows:

$$T_2 = \begin{cases} T_1 & \text{if } \ddot{d_{ij}} = 0 \\ \frac{\dot{d}_{ij}}{\ddot{d_{ij}}} & \text{if } \Delta < 0 \\ \min(\frac{-\dot{d}_{ij} \pm \sqrt{\Delta}}{\ddot{d_{ij}}}) & \text{if } \min(\frac{-\dot{d}_{ij} \pm \sqrt{\Delta}}{\ddot{d_{ij}}}) \geq 0 \\ \max(\frac{-\dot{d}_{ij} \pm \sqrt{\Delta}}{\ddot{d_{ij}}}) & \text{if } \min(\frac{-\dot{d}_{ij} \pm \sqrt{\Delta}}{\ddot{d_{ij}}}) < 0 \end{cases} \tag{4.3}$$

If the acceleration term $\ddot{d_{ij}}$, is zero it reverts to the first order case and $T_2 = T_1$. When Δ is negative, T_2 is defined as the time of closest approach as there are no real roots. There would be two roots when Δ is zero or positive. In the case when the roots are both positive, we take the lower value as it is the earliest time that the vehicles will collide. If one root is positive and the other negative, we take the positive value as it represents a collision in the future, which is what we are

interested in predicting. When both roots are negative, we take the root with the higher value. Moreover, negative values of TTC indicate that there is no risk of collision, hence, they should all be treated the same.

Looming

The drawback with MTTC is that it assumes that the vehicles are on the same collision course which might not always be the case. Thus, looming, as introduced in Ward et al. (2015), is used to check whether the vehicles actually reach the point of intersection at the same time or if they simply pass one before another. To calculate looming, seven test points are chosen on the vehicle as shown in Figure 4.4. The loom points are biased to the front of the vehicle as predicting the likelihood of collision with this part of the vehicle is more useful to the driver than the end of the vehicle. The linear velocity of the loom point ($\bar{\mathbf{v}}_\mathbf{i}$) is calculated as follows:

$$\bar{\mathbf{v}}_\mathbf{i} = \mathbf{v}_\mathbf{i} + (\mathbf{p}_\mathbf{i} - \mathbf{p}_\mathbf{c}) \times \omega_\mathbf{i} \tag{4.4}$$

Where $\mathbf{v}_\mathbf{i}$ is the ego vehicle velocity, $\mathbf{p}_\mathbf{i} - \mathbf{p}_\mathbf{c}$ is the displacement of the loom point ($\mathbf{p}_\mathbf{i}$) from the vehicle center of rotation ($\mathbf{p}_\mathbf{c}$) and $\omega_\mathbf{i}$ is yaw rate of the vehicle. The vector sum of vehicle velocity and the linear velocity due to the yaw of the vehicle about its centre gives the linear velocity of the loom point. Thus, the loom rate (angular velocity of the loom point) is calculated as follows:

$$\dot{\theta} = \frac{(\mathbf{p}_\mathbf{j} - \mathbf{p}_\mathbf{i}) \times \bar{\mathbf{v}}_\mathbf{i} + (\mathbf{p}_\mathbf{j} - \mathbf{p}_\mathbf{i}) \times \mathbf{v}_\mathbf{i}}{\|\mathbf{p}_\mathbf{i} - \mathbf{p}_\mathbf{j}\|^\mathbf{2}} \tag{4.5}$$

where $\mathbf{p}_\mathbf{j}$ is the vector position of the agent.

This gives rise to fourteen loom rates corresponding to the left loom rates (named alpha1 through alpha7) and the right loom rates (named beta1 through beta7) of the seven loom points. This calculation helps to determine if the vehicles are on a collision course.

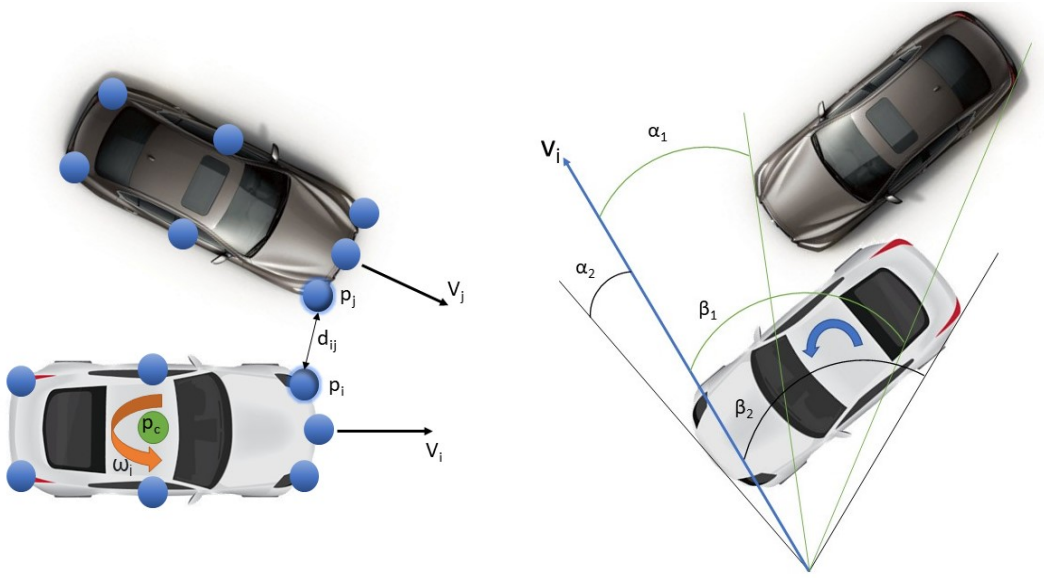


Figure 4.4: The blue circles denote the loom points. The right diagram shows the first four angles corresponding to the loom rates. In the right diagram, the rightmost point of the object is moving clockwise relative to the observer and the leftmost point is moving anticlockwise. Thus, the observer's field of vision is filled increasingly by the object and the object is looming.

4.5.2 Problem Statement

Given the following parameters: 14 loom rates, target destination, relative distance, relative velocities and acceleration, relative yaw, agents type, TTC_1 and TTC_2, design a collision risk prediction algorithm/model $\mathcal{M}$ and generate a factual and a counterfactual explanation for each risk prediction (r).

4.5.3 Algorithm Design

We adopt the proposed tree-based representation which offers better transparency. We propose the use of already established tree-based models, e.g., decision trees that have increased transparency, to learn the collision risk metric (in Section 4.5.1). In particular, tree-based models avail the opportunity to inspect decision boundaries and split criteria at every decision node. This makes the generation of intelligible natural language explanations possible. Similar approaches have been used by Stepin et al. (2021) for explaining fuzzy systems. We also devised a simple technique for choosing a representative tree from a forest to base our explanations on when the

model under consideration is a random forest model.

Algorithm 1: Tree-based Factual Explanation (FE1)

Algorithm 1: Tree-based Factual Explanation

Input: tree model $\mathcal{M}$, input vector V, predicted risk r
Output: intelligible textual factual explanation

```
1  causes ⟵ ∅
2  if isEnsemble(M) then
3      r_path ⟵ ∅
4      if isClassification(M) then
5          M ⟵ obtainModeTrees(M, V)
6          for m ∈ M do
7              r_path ⟵ r_path ∪ obtainDecPath(m, V)
8          M ⟵ obtainTreebyFactoringPaths(M, r_path)
9      else if isRegression(M) then
10         M ⟵ obtainMedianTree(M, V)
11 r_path ⟵ obtainDecPath(M, V)
12 causes ⟵ mergeInequilities(r_path)
13 return decode(r, causes)
```

In the tree-based factual explanation algorithm (FE1, i.e., Algorithm 1), we first check whether $\mathcal{M}$ is a random forest model (line 2). We further check whether $\mathcal{M}$ is a classification model or regression model (line 4). If $\mathcal{M}$ is a random forest classifier, we obtain all the trees in the forest that predicted the action r (line 5). Subsequently, the paths from the root to the r node of the resulting trees are obtained and the tree that contains the most re-occurring features across these trees is selected (lines 6 - 8). If the model is a random forest regressor, we obtain the tree that predicted the closest value to the median of all predicted r (lines 9 - 10). The decision path for the resulting classification or regression tree is obtained and the feature conditions along this path are merged to have a single decision boundary for each feature (lines 11 - 12) as there can be many split conditions for each feature. We refer to these merged conditions as causes. For example, if we have two conditions in a path: $\{\{`Feature1' < 50\}, \{`Feature1' < 10\}\}$, the merged path will be $\{`Feature1' < 10\}$. The decode function (line 13) provides a textual factual explanation based on the supplied causes.

If a counterfactual explanation is desired, a corresponding counterfactual explanation is generated using CFE1 (Algorithm 2).

Algorithm 2: Tree-based Counterfactual Explanations (CFE1)

Algorithm 2: Tree-based Counterfactual Explanation

Input: tree model $\mathcal{M}$, input vector V, predicted collision risk r, expected counterfactual output $r^{'}$

Output: intelligible textual counterfactual explanation

1 $r_{cfpath} \longleftarrow \emptyset$
2 $n_a \longleftarrow \emptyset$
3 **if** $r^{'} \equiv \emptyset$ **then**
4 $\quad r^{'} \longleftarrow findClosestCFSibling(\mathcal{M}, V, r)$
5 $n_r, r_{cfpath} \longleftarrow lowestCommonAncestor(\mathcal{M}, V, r, r^{'})$
6 $r_{cfpath}[n_r] \longleftarrow \neg r_{cfpath}[n_r]$
7 $conditions \longleftarrow mergeInequilities(r_{cfpath})$
8 **return** decode($r^{'}$, $conditions$, r_{cfpath})

To construct a counterfactual explanation, CFE1 (Algorithm 2) first checks whether a desired counterfactual output ($r^{'}$) was provided (line 3). When not provided, it finds the closest sibling node ($r^{'}$) to the leaf node (r) subject to the constraint that $r \neq r^{'}$ (line 4).

When $r^{'}$ is provided, the algorithm only finds the lowest common ancestor of r and $r^{'}$ obtaining the counterfactual conditions in the process for explanations (line 5). The condition at node n_r is negated and r_{cfpath} is updated to contain only counterfactual conditions (line 6).

4.6 Case Study 2: AV Action Explanations

AVs estimate risks before they commit to an action. Hence, having described how explainers could be designed for collision risk models, it is as well necessary to investigate how we can design explainers for AV navigation models. While explanations for a collision risk prediction based on vehicle dynamics are helpful for developers when they use a tree-based model, highly intelligible explanations about AV navigation actions based on what is observed in the environment are essential

for lay users. Hence, we describe methods for generating more passenger-friendly explanations for AV navigation actions.

4.6.1 Preliminaries

First, we consider an autonomous vehicle $a_v \in \mathcal{A}$ as a special type of agent (i.e., a driverless car) in a shared environment. $\mathcal{A}$ is a set of agents of different classes. Along with a_v, other agents (say $a_i \in \mathcal{A}$) also exist in the shared environment.

Each a (including a_v and a_i) has information $Y \subseteq \mathcal{Y}$ indicating its class (C_a), action ($\mathcal{X}_a$) and position ($\mathcal{P}_a$) at a given time $t \in \mathcal{T}$. So, $Y_t : Y_t = \{C_a, \mathcal{X}_a, \mathcal{P}_a\}$. Time is a real number $\mathcal{T} = \mathbb{R}^+$. a_v needs to observe other agents to plan its trajectory while respecting certain constraints (e.g. road rules and restrictions). AV's planned trajectory obtained at time $t \in \mathcal{T}$ is denoted as $\xi_{a_v}(t)$. $\mathcal{M}$ denotes a tree-based model that receives feature vector V as input and predicts the AV's current action $\mathcal{X}_{a_v}$. V may contain information Y_t of other relevant agents in the shared environment, and the AV's planned trajectory $\xi_{a_v}(t)$.

We represent actions ($\mathcal{X}_a$) and trajectories $\xi(t)$ as high-level SatNaV commands e.g., move, stop, lane change, among others. $\mathcal{G}$ is used to denote the set of constraints for generating explanations—especially counterfactual explanations—for the AV's actions.

4.6.2 Problem Statement

Generally, our goal is to predict $\mathcal{X}_{a_v}$ given V and generate posthoc intelligible explanations E for the prediction. For E to be intelligible, it should reference the influential features in V using clear high-level natural language semantics. We want to be able to:

- generate intelligible factual (or Why) explanations;
- select the relevant causes for action amidst different competing causes while generating factual explanations;

- generate intelligible counterfactual (or *What If*) explanations while respecting some constraints $\mathcal{G}$. In our case, constraints are restrictions in the form of input features whose attributes should not be modified when generating a counterfactual explanation;

- provides a sense of confidence for the explanation generated.

4.6.3 Algorithm Design

First, we must intelligibly describe what the sensors perceive from the driving workspace ($\mathcal{W}$) in order to be able to make meaningful explanations (E). More specifically, observations are a combination of other static and dynamic agents on the road ($\mathcal{A}$), the positions of each of the agents ($\mathcal{P}$), and the current action being performed by the agent (or the state of the agent) ($\mathcal{X}$), similar to the definition in Singh et al. (2021). For road traffic observations description, we propose a systematic approach that segments traffic scenes into different compartments and then describes the agents in each of these compartments from the perspective of the AV. For example, we partition lanes into different segments and name them based on the relative headings/direction of agents with respect to the AV. Below, we describe the lane positions resulting from this approach.

Lane positions

First, we refer to the lane on which the AV is on as the *ego lane*. Following the UK driving convention, agents moving in the opposite direction to the AV would definitely be on a lane to the right side of the AV. We refer to this lane as the *incoming lane*. This lane might sometimes be adjacent to the ego lane. Every other vehicle lane on which traffic flow in the same direction to the AV's direction is referred to as an *outgoing lane*. Depending on the road type and the lane position of the AV, the outgoing lane could either be on the left or right of the AV. There are also other vehicle drivable road positions, this includes junctions and crosswalks. Other road positions that are not vehicle drivable include the left and

right pavements, incoming cyclist lane and outgoing cyclist lane, parking lot, and bus stop. We describe a systematic agents categorisation scheme in the next section.

Agents

Agents are the different object types in a traffic scene. An agent could be active (that is, has a changing state) or passive (no state, e.g., houses). Each active agent belongs to a class/type (C_a) e.g., truck, cyclist. Active agents (e.g., pedestrians, cyclists, traffic lights, and other vehicles) tend to influence the AV's behaviour the most. Therefore, we focus on the active agents.

Actions

Actions are defined based on the lateral and longitudinal movements of dynamic vehicle agents on the road relative to the AV. Longitudinal actions include: moving away, moving towards, stopped. Lateral actions include left and right turns, left and right lane changes, crossing from left, and crossing from right. Pedestrians' actions are random and could be any of the actions in the two categories we have listed. Static agents such as traffic lights have states. A traffic light's state could be any of red, amber or green.

Thus, observations could be simply described by a combination of the agent's class, action and position.

$$\mathcal{O} :=< \mathcal{C}, \mathcal{X}, \mathcal{P} > \tag{4.6}$$

These high-level semantics—including how much of an influence they are on the AV's action—should be included in the annotations of driving datasets in order to facilitate the generation of human-understandable or intelligible explanations. Additional layers of descriptions can be added, e.g., influence tags to reflect how much of an influence an agent is on the AV's decision. Driving commentary from a human driver could be added to augment datasets. We provide an example data schema for data annotation, augmented with driving commentary. See Figure 4.5.

Figure 4.5: An example dataset annotation schema for explanations support. This example assumes that there is a human driver controlling the ego vehicle. The activities of the driver were also noted, e.g., commenting on a scene. The influence tag and the mention tags are introduced. The influence tag indicates whether the agent in question is the primary influence on the AV's action. The mention tag is used to indicate that the agent was mentioned in the driver's comment.

Algorithm 3: Tree-based Factual Explanations with Entropy Estimation (FE2)

In FE2 (Algorithm 3), we pass as input a tree model $\mathcal{M}$ which has been trained on a dataset containing records of encoded information Y_t for different agents in a shared environment at time t. We also pass the AV action predicted by $\mathcal{M}$. $\mathcal{M}$ can be a classification or regression decision tree, and can also be an ensemble of such trees.

In the tree-based algorithm, we first obtain the Contextual importance (CI) values for the feature attributes in V (line 2). The CI values express the importance of the different feature attributes for a prediction. Apart from being important, we want to know the extent to which the attributes of the different input features are favourable (or not) for a prediction, this is referred to as contextual utility (Anjomshoae et al., 2021). We check whether $\mathcal{M}$ is a random forest model. We further check whether $\mathcal{M}$ is a classification model or regression model. If $\mathcal{M}$ is a random forest classifier, we obtain all the trees in the forest that predicted the action $\mathcal{X}_{a_v}$ (line 6). Subsequently, the paths from the root to the $\mathcal{X}_{a_v}$ node of the resulting trees are obtained and the tree that contains the most re-occurring features across these trees is selected (lines 7 - 9). If the model is a random forest regressor, we obtain the tree that predicted the closest value to the median of all predicted $\mathcal{X}_{a_v}$ (lines 10 - 11). The decision path for the resulting classification or regression tree is obtained

Algorithm 3: Tree-based Factual Explanation with Entropy Estimation

Input: tree model $\mathcal{M}$, input vector V, predicted AV's action $\mathcal{X}_{a_v}$
Output: intelligible textual factual explanation

```
1  causes ⟵ ∅
2  ci ⟵ obtainCI(M, V)
3  if isEnsemble(M) then
4      X_path ⟵ ∅
5      if isClassification(M) then
6          M ⟵ obtainModeTrees(M, V)
7          for m ∈ M do
8              X_path ⟵ X_path ∪ obtainDecPath(m, V)
9          M ⟵ obtainTreebyFactoringPaths(M, X_path)
10     else if isRegression(M) then
11         M ⟵ obtainMedianTree(M, V)
12 X_path ⟵ obtainDecPath(M, V)
13 causes ⟵ mergeInequilities(X_path)
14 selected ← obtainRelevantCauses(ci, causes)
15 entropy ← Entropy(M, V)
16 return decode(X_{a_v}, selected, X_path, entropy)
```

and the feature conditions along this path are merged to have a single decision boundary for each feature (lines 12 - 13) as there can be many split conditions for each feature. We refer to these merged conditions as causes. For example, if we have two conditions in a path: $\{\{`Feature1' < 50\}, \{`Feature1' < 10\}\}$, the merged path will be $\{`Feature1' < 10\}$.

We then look for the causes whose features have high positive CIs (line 14). We do this by obtaining the cause with the maximum CI, and then adding more causes if the percentage difference between their CIs and the maximum CI is less than a threshold (we used a threshold of 50%).

We estimate the confidence of the model for each prediction by computing the information entropy of the training sample distribution in the leaf node of the decision path (line 15). This is obtained by:

$$H = \sum_{i} -p_i log_2 p_i \tag{4.7}$$

p_i is the probability of belonging to the i-th class. Low entropy reflects high confidence. While estimating CI (in line 2), we explored two methods—e.g., local

feature increments (Palczewska et al., 2013) and Tree SHAP (Lundberg & Lee, 2017)—that could provide this information.

A local feature increment for feature f denotes the difference in the probability of belonging to a class (say C_i) between the parent node (n_p) and the child node (n_c), given that f is the splitting feature in the parent node.

$$LI^f_{n_c} = \begin{cases} D_{n_c} - D_{n_p} & \text{if the split in the parent is performed over the feature } f, \\ 0 & otherwise \end{cases} \tag{4.8}$$

D_n is used to denote the fraction of the training instances in a node n that belongs to class C_i. The contextual importance of a feature f is the sum of $LI^f_{n_c}$ over all nodes on the path of the input instance from the root node to a terminal node.

Tree SHAP estimates the shapely values of each feature i from $1, ..., N$. The fundamental procedure is given as:

1. generate all subsets S of the set $F = \{1, ..., N \setminus \{i\}\}$

2. for each $S \subseteq F \setminus \{i\}$ estimate the contribution of feature i as:

$$CT\{i|S\} \;=\; \mathcal{M}(S \cup \{i\}) - \mathcal{M}(S) \tag{4.9}$$

3. compute the SHAP value according to:

$$\phi_i := \frac{1}{N} \sum_{S \subseteq F \setminus \{i\}} \binom{N-1}{|S|}^{-1} CT(i|S) \tag{4.10}$$

We chose Tree SHAP algorithm as it led to a better performance in terms of the BiLingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002) used to assess the resulting textual explanations (see results in Appendix A).

High positive SHAP values indicate high importance and utility, while very low negative SHAP values indicate high importance but low utility. The combination of CIs and the conditions along the decision path allows for the provision of more specific and intelligible explanations. The decode function (line 16) provides the human-understandable textual explanation based on the supplied values.

Algorithm 4: Tree-based Counterfactual Explanation (CFE2)

If a counterfactual explanation is desired, a corresponding counterfactual explanation is generated using CFE2 (Algorithm 4).

Algorithm 4: Tree-based Counterfactual Explanation

Input: tree model $\mathcal{M}$, input vector V, predicted AV's action $\mathcal{X}_{a_v}$, expected counterfactual output $\mathcal{X}'_{a_v}$, constraints $\mathcal{G}$

Output: intelligible textual counterfactual explanation

1 $\mathcal{X}_{cfpath} \longleftarrow \emptyset$
2 $n_a \longleftarrow \emptyset$
3 **if** $\mathcal{X}'_{a_v} \equiv \emptyset$ **then**
4 $\quad \mathcal{X}'_{a_v} \longleftarrow findClosestCFSibling(\mathcal{M}, V, \mathcal{X}_{a_v}, \mathcal{G})$
5 $n_a, \mathcal{X}_{cfpath} \longleftarrow lowestCommonAncestor(\mathcal{M}, V, \mathcal{X}_{a_v}, \mathcal{X}'_{a_v}, \mathcal{G})$
6 $\mathcal{X}_{cfpath}[n_a] \longleftarrow \neg\mathcal{X}_{cfpath}[n_a]$
7 $conditions \longleftarrow mergeInequilities(\mathcal{X}_{cfpath})$
8 $entropy \leftarrow Entropy(\mathcal{M}, V, \mathcal{X}_{a_v})$
9 **return** decode($\mathcal{X}'_{a_v}$, $conditions$, $\mathcal{X}_{cfpath}$, $entropy$)

To construct a counterfactual explanation, CFE2 (Algorithm 4) first checks whether a desired counterfactual output ($\mathcal{X}'_{a_v}$) was provided (line 3). When not provided, it finds the closest sibling node ($\mathcal{X}'_{a_v}$) to the leaf node ($\mathcal{X}_{a_v}$) subject to the constraint that $\mathcal{X}_{a_v} \neq \mathcal{X}'_{a_v}$ (lines 3 - 4). Another type of constraint used is such that restricts the modification of a feature attribute while searching for counterfactual candidates.

When $\mathcal{X}'_{a_v}$ is provided, the algorithm only finds the lowest common ancestor of $\mathcal{X}_{a_v}$ and $\mathcal{X}'_{a_v}$ obtaining the counterfactual conditions in the process (line 5). The condition at node n_a is negated and $\mathcal{X}_{cfpath}$ is updated to only contain counterfactual conditions (line 6).

Figure 4.6 illustrates how the tree explainers work.

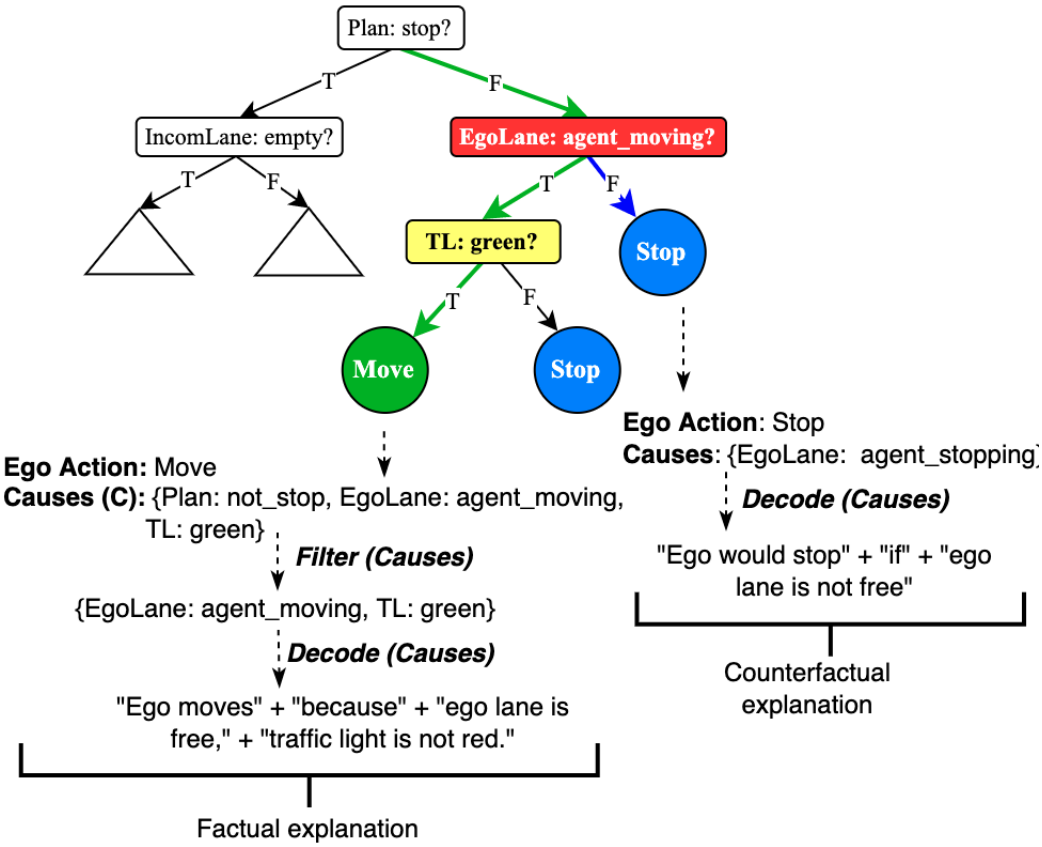


Figure 4.6: Explanation generation process. The green edges the model's decision path for a move action prediction. The conditions/causes on the decision path are filtered based on CI values. The decode function provides the natural language texts based on a predefined mapping of node conditions to English phrases. The blue nodes are the counterfactual candidates. The yellow node has a feature condition that is non-modifiable based on a set constraint. The closest sibling node to the predicted move action is the blue node below the yellow node, but because there is a constraint on the feature in the yellow node, we move a level up the tree to find the next closest sibling which is the blue node below the red node. The selected counterfactual candidate is the blue path from the red node to this new sibling node (the rightmost blue node). The condition in the red node is negated and the resulting list of conditions/causes is decoded to form a natural language counterfactual explanation.

4.7 Conclusion

In this chapter, we have taken a bottom-up approach to disambiguate explainability in autonomous driving. We did this by (i) proposing a conceptual framework for explainable AVs, (ii) proposing different autonomous driving data representations and proposing algorithms for generating posthoc factual (*Why*) explanations and counterfactual (*What If*) explanations. These explanations meet specific requirements such as intelligibility to humans, selectivity, and the provision of a sense of confidence through entropy estimation. We looked at two case studies: collision risk explanation and AV action explanation. One limitation of the work in this chapter is that the proposed algorithms rely on the outputs from a proxy tree-based model. This makes the resulting explanations 'approximate explanations' in that we cannot completely assure their faithfulness.

In the next chapter, we apply the proposed algorithms in different experiments and report their performances against ground truth explanations.

5

Explanation Generation: Experimental Results

Contents

5.1 Introduction

In this chapter, we describe the experiments conducted to substantiate and demonstrate the theories, structures and algorithms presented in Chapter 4. **This chapter addresses the research question regarding how the proposed explanation**

generation algorithms can be effectively applied to generate the types of explanations explored in Chapter 3. We apply FE1 (Algorithm 1) and CFE1 (Algorithm 2) on the Lyft Level5 dataset to explain a collision risk prediction model. We then introduce our new driving dataset (SAX dataset) which we have annotated following the data representations and structures discussed in Chapter 4, in order to support the development of explainable autonomous driving models. The Lyft Level5 and the SAX datasets were both obtained by an ego vehicle, driven along different vehicle routes. We applied FE2 (Algorithm 3) and CFE2 (Algorithm 4) on the new SAX dataset to explain a navigation action prediction model, a surrogate model that learns the ego vehicle's behaviour. We present the results from the experiments conducted for the collision risk and ego vehicle navigation cases.

5.2 Case Study 1: Explaining Collision Risk

We quantified the risk posed to an autonomous vehicle by other road users (agents) present in the environment at any given instant. This is further used to predict the risk of collision at various time horizons into the future. Figure 5.1 describes a scene from the Lyft Level5 dataset with predicted collision risk values. The predicted risks of collision and causes of such risks could be communicated as explanations to serve as timely warnings for impending dangers around the AV. Moreover, these explanations could help developers improve their tree-based collision risk models.

5.2.1 Experiment

We conducted an experiment to assess and explain collision risk based on the planar TTC with the looming metric defined in Chapter 4. We extracted relevant features from the Lyft dataset to learn a risk model. We further applied FE1 (Algorithm 1) and CFE1 (Algorithm 2) to generate explanations for the model's predictions. The first aspect of this section describes the setup for collision risk assessment, and the later section describes how explanations are generated for collision risk predictions.

Lyft 'level5 Dataset

In order to evaluate the efficacy of the proposed algorithms, we utilised the Lyft Level 5 Prediction dataset (Houston et al., 2020). The dataset is primarily composed of a set of scenes collected across 1,118 hours of autonomous driving activity from 20 different vehicles. These scenes are accompanied by a manually constructed semantic map detailing the road network, as well as a metadata file. The metadata describes part of the transformation from the semantic map frames—based upon a geodetic datum—to the world frame used for the scene data. The scene data is structured as follows: the top level, which consists of a series of scenes. Each scene is $\sim 25\,s$ and is composed of ~ 250 frames sampled at 1 *Hz*. Each frame contains a timestamp, translation and rotation values for the ego vehicle, and the relevant agent objects. Agent objects do not persist between frames, and instead, use a tracking id to identify the same entity between frames. For our experiments, the first 300 scenes were used to generate feature vectors, of which 20% were used for testing with five-fold cross-validation.

Feature Extraction

From the Lyft Level5 dataset, we extracted the vehicles' physical quantities described in Chapter 3. The relative distance between the ego vehicle and the agent was calculated by extracting the ego and the agent's current positions and taking the L^2 norm of their difference. The agent's velocity and the ego's velocity were calculated by iterating through the frames and averaging their changing position over time. Similarly, acceleration was calculated by averaging instantaneous velocity over time. The relative velocity and the relative acceleration were calculated by obtaining the difference of each related pair of the two physical quantities, followed by an L^2 norm. The angular velocity of the ego was obtained by averaging its changing yaw over time. The relative yaw is the difference between the yaw of the agent and the ego. The target position of the ego vehicle is also included in the feature vector as it gives a sense of the direction the ego vehicle aims to move towards.

Feature Vector Generation

The information from Section 5.2.1 was combined to generate a feature vector which served as input for our machine learning-based collision risk assessment method. The information includes the following: T_1, T_2, the fourteen loom rates, relative distance, ego and agent velocity, relative velocity, agent and ego acceleration, relative acceleration, the angular velocity of the ego, the target destination of the ego in both the x and the y direction, the relative yaw and the type of agent. T_1 and T_2 were capped to 30 seconds as values above these are much larger than the horizon we wanted to predict the collision in and this might skew the data.

5.2.2 Ground Truth

To determine the ground truth (labels for our learning task), the actual relative distance between the ego vehicle and the agent at future time t was extracted.

Classification Labels

As applied in Ward et al. (2015), 10 metres was used as the threshold for classification. If the future relative distance was less than 10 metres, it was classified as high risk (risk flag = 1) and if it was greater than or equal to 10 metres, it was considered low risk (risk flag = 0). This is intuitive as the vehicles are not likely to collide when the distance between them is more than 10 metres.

Regression Ground Truth

While classification helps to distinguish between high and low-risk objects, it does not provide a rich calibration scale for collision risk. Hence, we decided to extend this to a regression problem. To obtain ground truth labels (risk scores), we sampled the probability of the actual distance from a one-sided positive Gaussian distribution with zero mean and a standard deviation of five. Two standard deviations correspond to 10 meters; thus, most of the data points were included.

Prediction Times

The ground truth labels for regression and classification were generated at 1, 3 and 5 seconds into the future. Reaction times vary greatly from person to person, and even for the same person it changes based on the time of the day, weather conditions and the landscape (Palazzi et al., 2018). A professional driver who is physically fit and trained in high-speed driving might have a reaction time of 0.2 seconds for a given situation, while the average driver may have a slower reaction time of 0.5 seconds, 0.8 seconds or even 1-second ("Managing a Slow Reaction Time While Driving", 2019). However, in cases where the human driver needs to override and take control, we need more time since human drivers in autonomous vehicles tend to be more disengaged in the task and more overconfident in the automation. They can have a weakened understanding of the operation and status of the automatic system, as well as that of the driving situation the car is in. In the long term, they could also lose the skills required to drive and operate the car safely (Demmel et al., 2019). Thus, even longer time horizons such as 3 and 5 seconds were included as explained in Figure 5.1.

Collision Risk Model Performance

Table 5.1 compares the results from different classification models while Table 5.2 compares the results from different regression models. While the difference between the decision tree and random forest models is not very evident in the classification case, they are very distinct in the regression case. Performance decreased as we increased the time horizon for our predictions. This is expected as it is difficult to make accurate predictions of a dynamic human-driven vehicle whose attributes can change in less than a second. For example, it is more likely that our assumption of constant acceleration or deceleration may hold true for $1\,s$ rather than for $5\,s$. Table 5.3 compares the regression model among different agent types. The 'count' column represents the number of instances of the particular class present in the dataset. The models perform the best on cars and the least on pedestrians. This is due to the relative ease in estimating the velocity and acceleration of vehicles

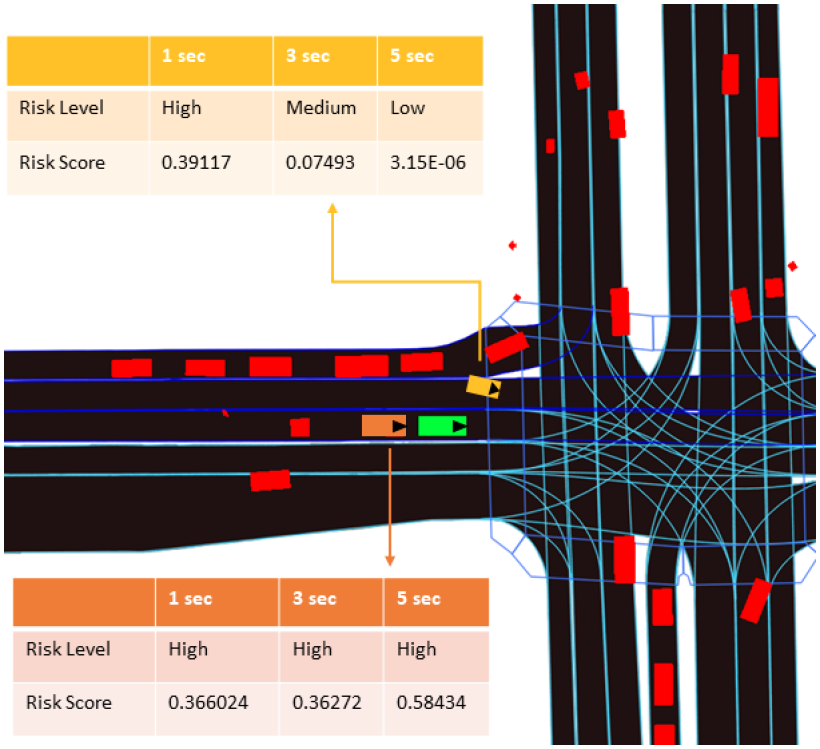


Figure 5.1: An instance of the Lyft Level5 dataset (Houston et al., 2020). The green rectangle represents the ego vehicle, the yellow rectangle is the agent whose risk value is predicted relative to the ego vehicle. The orange rectangle represents another different agent behind the ego vehicle. The tables show the ground truth risk prediction values for 1 second, 3 seconds, and 5 seconds. The black triangle points in the direction the vehicle is moving. In Section 5.2.2, we provide explanations for the risk predicted for the yellow agent.

compared to pedestrians. Moreover, as pedestrians are structurally different from cars (i.e., with lesser surface areas), the loom point method might not be quite accurate for estimating collision risk with pedestrians.

Table 5.1: Comparing Different Classification Models.DT: Decision Tree, RF: Random Forest

Time	Model	RMS Error	AUC	F_1 Score
1 sec	DT	0.313112	0.91	0.873064
	RF	0.280056	0.98	0.895877
3 sec	DT	0.285831	0.90	0.851776
	RF	0.291492	0.92	0.844609
5 sec	DT	0.313112	0.86	0.784305
	RF	0.291492	0.88	0.816935

Table 5.2: Comparing Different Regression Models

Time	Model	RMS Error	EVS	R^2 Score
1 sec	DT	0.089790	0.513486	0.509954
	RF	0.036018	0.921205	0.921146
3 sec	DT	0.006891	0.449383	0.447278
	RF	0.051660	0.786132	0.785953
5 sec	DT	0.091964	0.369128	0.369034
	RF	0.058059	0.7485211	0.748506

Table 5.3: Comparing Different Classes

Class	Count	RMS Error	EVS	R^2 Score
Car	519385	0.076093	0.754399	0.752412
Cycle	6688	0.053561	0.864631	0.735844
Ped	43182	0.127931	0.695486	0.638030

Explanation Algorithm Performance

We applied FE1 (Algorithm 1) and CFE1 (Algorithm 2) to explain the predictions of the risk collision model. Explanations 1, 2, and 3 below are sample explanations generated using the tree-based method for the predicted collision risk of the ego vehicle with the yellow agent in Figure 5.1. We generated natural language explanations ('why' and 'what-if') for the tree models' predictions.

Explanation 1: RandomForest Regressor, $1s$
Why: "The predicted risk for the provided agent's attributes is 0.4922 because important features such as 'beta6' has a value between $0.0\,rad\,s^{-1}$ and $16.0179\,rad\,s^{-1}$, 'agent_vel' was below $5.2209\,ms^{-1}$, 'ego_vel' was below $0.0001\,ms^{-1}$."
What-If (counterfactual inference): "To get the risk prediction below 0.3, the following conditions should be true: 'alpha6' should be greater than $0.0\,rad\,s^{-1}$, 'agent_vel' should be above $6.794\,ms^{-1}$."

Explanation 1 was generated for a $1s$ random forest regressor prediction for a particular feature vector (say X). The counterfactual explanation is also generated for risk values lower than 0.3. When 'agent_vel' was set to $7\,ms^{-1}$, a risk value of 0.2614 was obtained. Increasing the 'agent_vel' makes the agent move farther

ahead of the ego vehicle, thereby reducing collision risk.

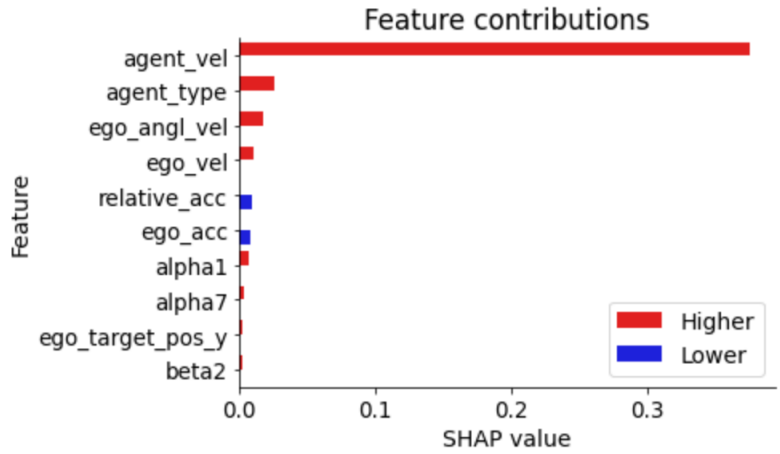
Explanation 2: RandomForest Regressor, $5s$
Why: "The predicted risk for the provided agent's attributes is 0.3853 because important feature such as 'beta2' was above -1.105e-05 $rad\,s^{-1}$, 'agent_vel' was below 5.1108 ms^{-1}, 'ego_target_pos_y' was below 0.6182 m."
What-if (counterfactual inference): "To get the risk prediction below 0.3, the following conditions should be true: 'ego_target_pos_y' should be greater than 0.6182 m."

Explanation 2 was generated for a $5s$ random forest regressor prediction for feature vector X. The counterfactual explanation was generated for a risk value lower than 0.3. When 'ego_target_pos_y' was set to 3, a risk value of 0.2754 was obtained. When the ego vehicle's target y is increased, the ego vehicle's destination is further south (where $(0, 0)$ is the topmost left corner) which makes its trajectory farther apart from the agent when the agent is heading east.

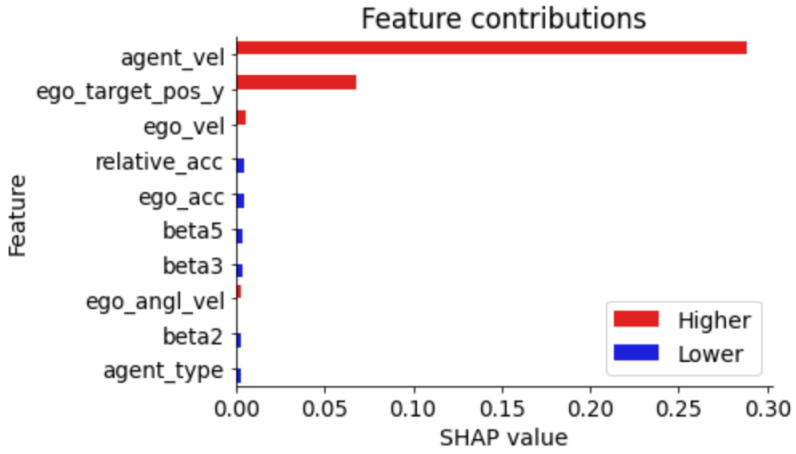
Explanation 3: RandomForest Classifier, $1s$
Why: "The provided agent was classified as 'high risk' because important feature such as 'alpha1' was below 1.6972 $rad\,s^{-1}$, 'alpha5' has a value between -180.2083 $rad\,s^{-1}$ and 0.0 $rad\,s^{-1}$, 'alpha7' was above -0.00046231 $rad\,s^{-1}$, 'beta1' was above -2.9e-07 $rad\,s^{-1}$, 'agent_vel' was below 23.5176 ms^{-1}, 'rel_yaw' was above -0.4258 rad."
What-if (counterfactual inference): "The closest class to the prediction is 'low risk'. To classify this sample as low risk the following conditions should hold: 'agent_vel' should be greater than 23.5176 ms^{-1}."

Explanation 3 was generated for a $1s$ random forest classifier's prediction for feature vector X. An explanation on how to obtain a counterfactual output (low risk) was also provided.

One way to verify the correctness of an explainer is to compare its explanations with the explanations from a different explainer but with the same input and check for correlation (Sippy et al., 2020). We qualitatively compared the outputs from FE1 (Algorithm 1) and CFE1 (Algorithm 2) with that from Tree SHAP algorithm (Lundberg & Lee, 2017) and noticed a high level of agreement. The



(a) Random Forest Regressor, 1 sec prediction



(b) Random Forest Regressor, 5 secs prediction

Figure 5.2: Results from SHAP explainer. This is a plot indicating feature contributions for the example scene shown in Figure 5.1 (yellow agent). We show the 10 most contributing features to prediction based on SHAP values obtained from the Tree SHAP algorithm for $1s$ and $5s$ predictions. Both predictions are for feature vector X passed to the RandomForest Regressor models.

generated factual explanations highlighted similar salient features that were observed from Tree SHAP (see Figure 5.2).

The following limitations are associated with FE1 (Algorithm 1) and CFE1 (Algorithm 2):

- FE1 (Algorithm 1) does not support causes selection from many candidate causes. It outputs all causes in the explanations.
- CFE1 (Algorithm 2) does not support the introduction of constraints when generating counterfactual explanations.

- Both algorithms do not provide any confidence measure of the model's prediction.
- Explanations generated do not provide human-understandable texts as the causes are numerical descriptions of the feature vectors and their decision boundaries in the tree.

The explanations are useful for developers and engineers, as they can help them to identify the most influential risk features from learnt risk assessment models. Moreover, through counterfactual inference/explanations, the algorithm can provide explanations that describe how risk features could be manipulated to decrease the overall risk in safety-critical driving scenarios. Although CFE1 (Algorithm 2) does not support the introduction of constraints while generating counterfactual explanations, a desired counterfactual class can be explicitly set for a classification task. For a regression task, the desired range in which the counterfactual outcome must fall can be set while generating a counterfactual explanation.

5.3 Case Study 2: Explaining Driving Actions

In the previous case study (Case Study 1), we demonstrated how explanations could be generated from a tree-based model for collision risk predictions. In the current case study (Case Study 2), we describe our procedure for generating more intelligible approximate natural language explanations for navigation decisions in autonomous driving. First, we describe the data collection and annotation procedure of the SAX dataset; a dataset useful for training a navigation action prediction model. We then describe an experiment in which we trained the action prediction model and assessed the performance of our explainer algorithms FE2 (Algorithm 3) and CFE2 (Algorithm 4) on the trained model. See an overview of the method in Figure 5.3

5.3.1 The SAX Dataset

An important aspect of safe autonomous navigation is the detection and tracking of agents in the environment where the AV operates (Caesar et al., 2020). To

achieve safe navigation, multiple sensors and machine learning-based detection and tracking algorithms are being deployed for better scene understanding and planning. Lately, the dependence on machine learning-based scene understanding approaches is seen to be on the rise owing to the record performance (Tan & Le, 2021) of machine learning algorithms on large datasets, especially images and videos. This increasing dependence on machine learning-based scene understanding approaches drives the need for larger and richer datasets for benchmarking and further research. A few of such datasets already exist, for example, the Lyft Level 5 dataset (Houston et al., 2020), NuScene dataset (Caesar et al., 2020), the Waymo open motion dataset (Ettinger et al., 2021), among others. However, these datasets are created without explainability support in mind.

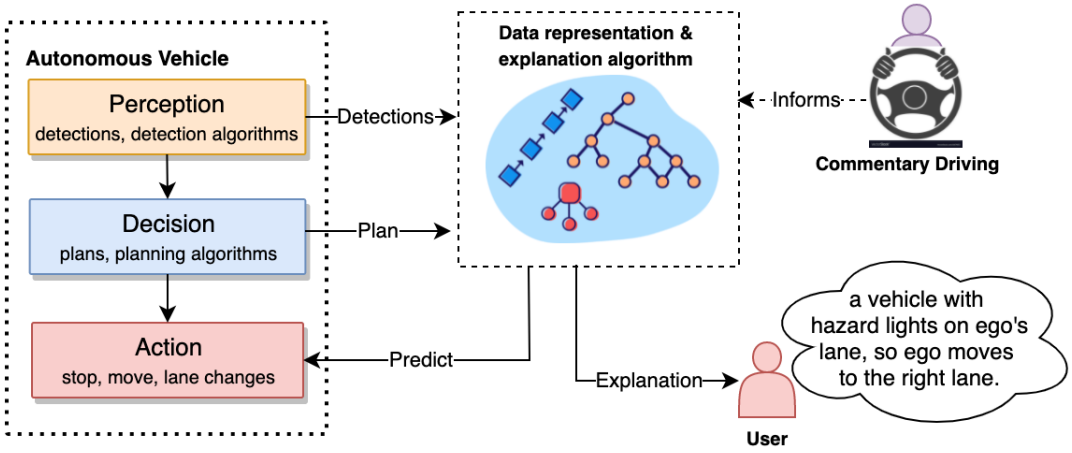


Figure 5.3: From commentary driving, requirements for explanations were gathered to inform the design of two explanation algorithms. The algorithm receives input data from the different autonomous driving operations, provides a structured representation, and generates intelligible explanations to stakeholders.

With the push towards explainability, it is critical that scene understanding and planning processes in AVs are transparent. Though different datasets—mostly vision-based—for autonomous driving have been introduced, most of them were not created with explainability in mind. Hence, they are more accessible for the development of deep vision-based models. To address this gap, we released the SAX dataset which provides richer information for explainable model development compared to the few existing explanation focused datasets, such as the ROad dataset (Singh et al., 2021) and the BDD-X dataset (J. Kim et al., 2018).

Similar to the ROad dataset (Singh et al., 2021), the SAX dataset provides semantically meaningful concepts for agents' intention prediction and planning. It extends the ROad dataset and the BDD-X dataset to become the first dataset to provide more structured information—through commentary driving—to support the development of explainable driving models. Generally, the information available in the SAX dataset includes:

- RGB image sequences from the dash cameras;
- multilabel annotations that provide rich semantic information about the ego vehicle and the other agents in the environment;
- internal states data from CAN logs;
- driving commentary provided by an expert driving instructor;
- structured explanations that provide clear causes for actions.

In the next section, we describe the methodology that was used to collect the SAX dataset.

5.3.2 Data Collection

Having obtained the necessary permissions from the University of Oxford's research ethics committee, we deployed an ego-vehicle to an urban environment with an expert driving instructor who drove the vehicle via different routes in London to collect the data necessary for our study. We asked the driver to provide commentaries while he drove. The ego-vehicle was fitted with different sensors among which were: microphones, cameras, and radar. The in-cabin microphone recorded the commentaries, while the external cameras recorded the environment scenes.

5.3.3 Driving Commentary Analysis

In order to understand the driving instructor's commentary/explanations style, we selected a representative 3-hour length video for video analysis. The video had different scenes with a series of navigation actions, such as 87 stops, 31 right lane changes (RLC), 29 left lane changes (LLC), and so many straight move actions. We ensured that all comments of the driver related to these actions were part of the collected video instances. We watched the video instances repeatedly for the four actions, and we discovered that the driver in most cases made explanations in the following order: (i) announce observations, (ii) announce plans (with or without reference to road rules), and (iii) make general remarks.

(i) Announce observations: The driver first announces an observation within his limit points. The limit point is the farthest point along a road to which a driver has a clear and uninterrupted view of the road surface. When there are many observations, he selects and focuses on the most interesting and relevant ones.

When describing an observation, the driver mentions the type of agent, the current action of the agent, and the road position of the agent.

(ii) Announce plans: Following the announcement of observations, the driver announces his planned action, and in some cases, how the plan respects the relevant road rule. In some cases, he states his plan during or after execution.

(iii) General remarks: When the driver completes the execution of his initial plan and the vehicle is steadily moving, he starts making a general comment about special observations in his view, e.g., road topology, special architectural designs, and pedestrians' movement on the sidewalks.

Excerpt from the description of an observation:

Driver (Video 3: 09:45 – 09:53): 'We've got a vehicle on the left part indicating so lights are on so he's live, we'd leave space so if he does move out [paused] now we'll pull back into the lane'
Driver (Video 2: 24:14 – 24:21): 'Got a bus pulling up on the left hand

> side and we've got lots of bright lights so there's gonna be traffic coming up there.'

The amount of space an agent in front of the driver occupied affected the driver's decision. The scene became an interesting one to comment on.

> *Driver (Video 1: 24:52 – 25:13)*: 'I could see but I'm quite a big vehicle. And by the time I've sort of gone in and gone out he's probably cleared anyway...Again, if I were on a little sports car, I might do it. But this is just a little bit bigger, I've got to be a bit more cautious with it.'

In many instances, the driver made counterfactual inferences:

> *Driver (Video 1: 06:55 – 06:59):* 'If I tried to go around the bus, I'd have been strained to the back of a vehicle that was already there.'

Generally, the most common causes for the driver's stop actions were traffic lights, pedestrians crossing, and static traffic queues. The common reasons for lane changes were to drive around a parked vehicle, to overtake a vehicle, to move to a faster lane, or to return to its default lane.

5.3.4 Data Annotation Scheme

We annotated all the data collected from the field trial. In our annotation processes, we made the best effort to annotate different driving scenarios with a mix of objective criteria and subjective judgement. For example, we were only interested in annotating interesting events where the behaviour of the ego changed and/or scenarios that the driving instructor made comments about. We followed the scenario categorisation by Ramanishka et al. (2018), which we also applied in Chapter 3 (i.e., the goal-driven and stimulus-driven scenarios). In this context, goal-driven scenarios include; left and right turns, and moves. Stimulus-driven scenarios include; lane changes and stop actions due to slow traffic. In our annotation scheme, an event in a scenario comprises active agents, their actions and locations relative to the ego vehicle. We use the tuple $\mathcal{O} =< \mathcal{C}_a, \mathcal{X}_a, \mathcal{P}_a, \mathcal{I}_a, \mathcal{D}_a, \mathcal{R}_a >$ to describe an observation. C_a is the class of the agent a, $\mathcal{X}_a$ is the action that agent a is performing, and $\mathcal{P}_a$ is the location of a relative to the ego vehicle. $\mathcal{I}_a$ indicates whether a is an influence on the action of the ego vehicle e.g., a red traffic light could

be an influence to the ego's stop action. $\mathcal{D}_a$ is used to indicate whether the driving instructor mentioned an agent a in his comment about an event. R_a is the road type where a is (e.g., single lane, double lane, and dual lane). $\mathcal{C}_a, \mathcal{X}_a, \mathcal{P}_a, \mathcal{I}_a, \mathcal{D}_a, \mathcal{R}_a$ are all members of a finite list compiled after inspecting the content of all of the videos collected (about 9.5 hours in length).

Based on these categorisations and definitions, we hired annotators who have driving knowledge to annotate the videos. We used the Microsoft VoTT annotation tool as it provides the ability to associate multiple labels for each bounding box; this aligns with our annotation scheme. Moreover, the tool provides an easy to use graphical interface and can run on most operating systems and on a web browser. It also allows copying bounding boxes across frames. We used the Otter.ai tool to transcribe the driver's comments for all videos. The transcribed comments were linked to their corresponding observations. We also provided a well-structured explanation for each event using the semantic information obtained from the annotations. Structured explanations take the form: $E =< observations/causes > + < egoaction >$. The influence and mention tags are particularly helpful in generating the causes for an ego's action. We also linked each event to its corresponding CAN bus data. To ensure consistency and error-free annotation, we had regular group meetings with all annotators to collectively inspect each annotator's work and corrected all errors. The final stage of quality assurance was the use of a computer programme to ensure that the annotated files had the right tags.

5.3.5 Comparison with Related Datasets

While our dataset was generated from only a 9.5-hour length video, it provides richer labels (spatial, temporal, and multiple labels for each bounding box) compared to the other datasets. It also provides more sensor information and richer explanations. See Table 5.4.

Table 5.4: Driving datasets with explanations support. BBox: Bounding box, Spat.: Spatial, Temp.: Temporal

Dataset	Hours	Annotations	Sensors	Explanation Type
BDD-X (J. Kim et al., 2018)	77	Spat., Temp.	RGB seq., vel., yaw, GPS	*Why explanation:* actions and descriptions (text)
BDD-OIA (Xu et al., 2020)	31.9	Spat., Temp.	RGB seq.	*Why explanation:* action and action inducing objects (text)
DoTA (Yao et al., 2020)	20.3	Spat., Temp.	RGB seq.	*What explanation:* anomaly identification (BBox)
CTA (You & Han, 2020)	9.5	Spat, Temp.	RGB seq.	*Why explanation:* accidents cause and effects (text)
HDD (Ramanishka et al., 2018)	104	Spat., Temp.	RGB seq., accl. & brake pedal, steering angle and steering speed, yaw rate	*Why explanations:* causes for stimulus-driven actions (text)
BDD-A Extended (Shen et al., 2020)	3	Temp.	RGB seq., speed, GPS	*Why explanation:* actions and causes (text); Gazemap (RGB seq); explanation necessity score (real number)
Road (Singh et al., 2021)	2.9	Spat., Temp. multi-label	RGB seq.	*What explanation:* agents' description (BBox)
Ours (SAX)	9.5	Spat., Temp., multi-label	RGB seq., CAN, GPS, IMU, accl. & brake pedals, steering angle, yaw and yaw rate	*Why explanation:* actions and causes (text), influence (BBox and text), driving commentary (text), ego next plan (text)

5.3.6 Dataset Statistics

Table 5.5 shows the different types of agents in our dataset (Agent column), the number of bounding boxes on agents across frames (# BBoxes column), and the number of labels per bounding box across the different agent classes (# Labels). The total number of bounding boxes and bounding box labels are 35,729 and 115,476 respectively.

For the rest of this chapter, we aim to demonstrate how intelligible natural language explanations can be generated for an AV's action, by describing observations (stating agent type, action, and position) and AV plans, as observed from an ego vehicle in our field study. We will show how explanations could be selective in the presence of competing causes. We will also demonstrate the

Table 5.5: SAX Dataset Annotation Statistic. BBoxes: Bounding Boxes

#	**Agent**	**# BBoxes**	**# Labels**
1	Ego	2,433	8,032
2	Car	18,098	59,641
3	Pedestrian	2,050	6,145
4	Van	3,167	10,514
5	Traffic Light	5,611	17,210
6	Cyclist	628	1,885
7	Bus	985	3,348
8	Truck	2,054	6,501
9	Motorbike	609	1,885
10	Emergency Vehicle	80	282
11	Others (e.g., roadblock)	8	23
	Total:	**35,729**	**115,476**

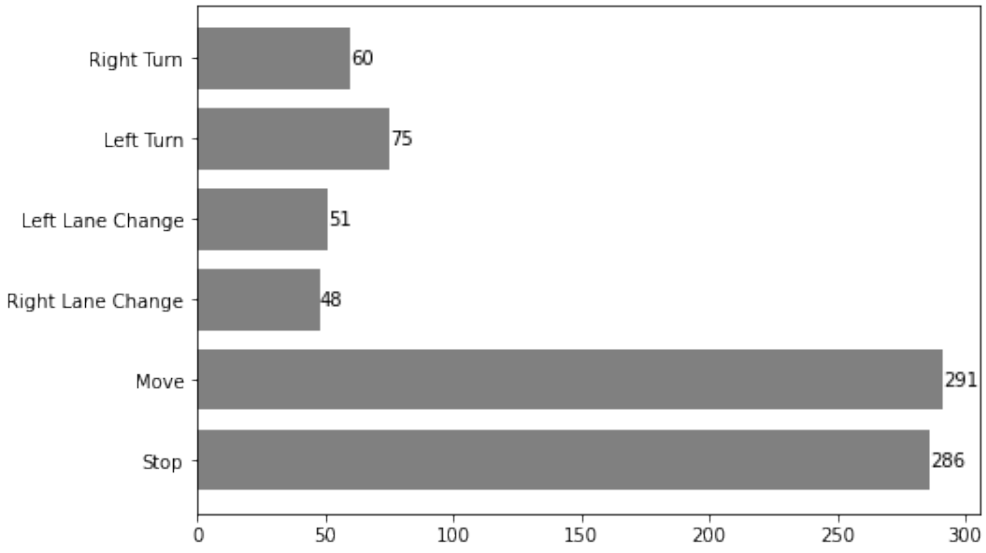


Figure 5.4: Ego's actions distribution in our dataset. There are a total of 811 actions. Each action is a sequence of RGB frames sampled at 6 frames per second from about 2 - 4 seconds of an action video. Stop and move actions dominate the dataset. This is mainly due to the road topology in London, the ego's goal, the presence of several traffic jams, and traffic lights. Each of these actions has accompanying explanations, driver commentary, ego's next plan, and different bounding boxes with multi-labels like agent class, action, position, influence, mention, and road type tags.

generation of counterfactual explanations with constraints. We reflect confidence in predictions using entropy scores.

5.3.7 Experiment

One way to explain the decisions of an AV with an end-to-end (black-box) architecture, is to develop a simpler model as a proxy to observe and learn the behaviour of the original black-box model after several input perturbations over time. This is similar to the perturbation explainable AI approaches where an interpretable model is used to learn the behaviour of a complex model by introducing small perturbations to inputs (Ribeiro et al., 2016). In our case, we assume an AV with a modular architecture which follows our conceptual framework. Hence, we gain access to a richer set of data—e.g., the perception and planning systems' outputs—to train our tree-based proxy model. This tree-based model then serves as one of the inputs to our explanation algorithm. This approach is limited, just as with other posthoc explanation methods. It is impracticable to assume that the proxy tree-based model is 100% faithful to the original model.

We fit a tree-based model on the SAX dataset to predict lateral and longitudinal navigation actions. The SAX dataset already provides us with perception and planning data. We further apply our explanation generation algorithms FE2 (Algorithm 3) and CFE2 (Algorithm 4) to provide posthoc explanations for the predicted navigation actions, which serve as approximate explanations for the ego vehicle's action. We qualify the explanations as 'approximate' because we can not guarantee that the generated explanations are always faithful. While our explainer has been trained on the SAX dataset of human driving behaviour, our explanation generation methods are transferable to actual AVs. In fact, AV models are often trained (or pre-trained) on datasets collected from ego vehicles that are driven by human drivers, e.g., in Acuna et al. (2021). The same driving rules and principles hold for both human and automated drivers. We argue that our explainer provides explanations based on these rules and principles.

In the next section, we explain how we created a subset of the SAX dataset to train model $\mathcal{M}$.

Feature Vector Extraction and Data Sampling

Observations in our experiment, among other interesting information, comprise agents of various classes $\mathcal{C}_{a_i}$ (e.g., vehicle, motorbike, pedestrian, traffic light), their actions $\mathcal{X}_{a_i}$ (e.g., moving, crossing), locations relative to the ego vehicle $\mathcal{P}_{a_i}$ (e.g., EgoLane, IncomLane, OutgoLane), plan/trajectory ξ_{a_i} (e.g., Move, Stop), influence on the ego vehicle (e.g., primary influence), and the driver's comment on the scene at the time. We generated the ground truth explanations for the ego's actions based on the influence tags that were added to the agents that influenced the ego's actions. We converted this semantic information into tabular forms where we have the following columns as features to train a tree model: EgoLane, IncomLane, OutgoLane, TL, EgoPlan, and EgoAction. EgoLane is the current lane that the ego vehicle travels on. The IncomLane is an adjacent lane on which traffic flows in the opposite direction to the ego vehicle's direction. The OutgoLane is an adjacent lane in which traffic flows in the same direction as the ego vehicle. Most of the relevant agents on the road fall within one of these lanes. TL is traffic light, and EgoPlan is the ego vehicle's high-level trajectory or simply the next action n-frames ahead. Agents with their corresponding actions were numerically encoded to serve as the attributes for these features. Where there was more than one agent on a lane, following the human driver approach from our field study, we selected the most dominant one. That is, the one that influenced the ego vehicle's decision the most; this is dependent on the size and proximity of the agent, this included pedestrians. Each record in our dataset represented a frame. We randomly sampled our training and test sets from this dataset. However, there were not as many lane change examples as stops and moves due to the rarity of such events in comparison with others.

The Explainer Model

We fitted a tree-based model (a random forest model, $\mathcal{M}$) on the curated tabular dataset. There were 2,755 sampled records from the original dataset, with 800 stop

instances, 900 move instances, 483 right lane change instances, and 572 left lane change instances of the ego vehicle. This sampling was done to reduce the effect of the very unbalanced nature of the dataset. The train/test split ratio was 80:20. The test set comprised 147 stop instances, 191 move instances, 105 right lane change instances, and 108 right lane change instances of the ego vehicle. Given a driving video, the tree-based model can be applied at each instance $t \in \mathcal{T}$ to predict the ego's action, which is subsequently explained using the described explanation algorithms.

The model yielded a test accuracy of 0.75 with higher performance for the stop and the move actions. This model $\mathcal{M}$ was used in FE2 (Algorithm 3) and CFE2 (Algorithm 4) to generate explanations. See more detail about the model's performance in Figure 5.5.

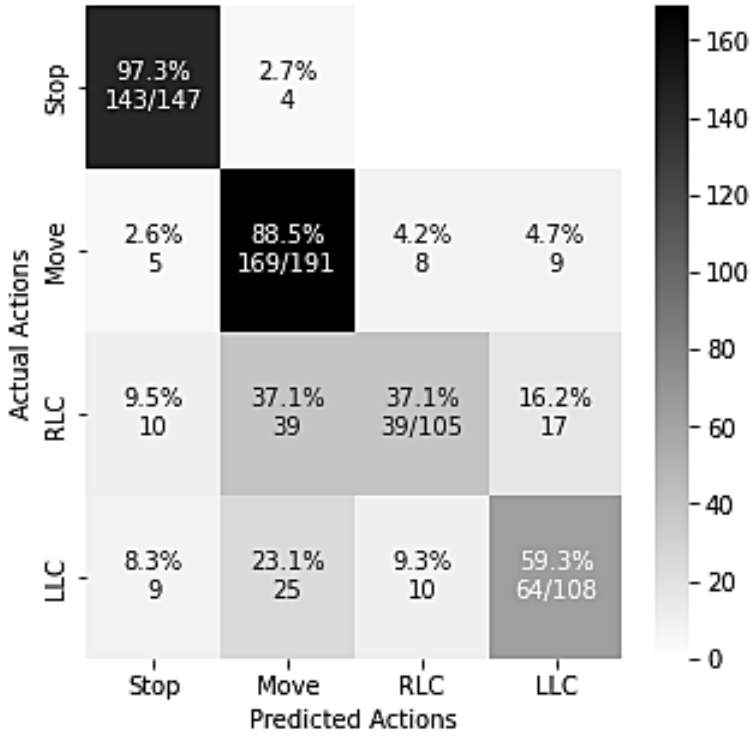


Figure 5.5: The model yielded a test accuracy of 0.75. with higher performance for stop and move actions.

5.3.8 Explanation Algorithm Performance

We first provide examples of scenarios in which FE2 (Algorithm 3) and CFE2 (Algorithm 4) were applied to generate explanations.

a.) Stop action b.) Move action c.) Right lane change action d.) Left lane change action

a.) Stop action explanation:
Factual Explanation: "Traffic light is not green on ego's lane, so ego stops" $S = 0.45$
Counterfactual Explanation: "If ego must move straight, the following should be happening: road is free on ego's lane; the traffic light is green on ego's lane" $S = 0.46$
b.) Move action explanation:
Factual Explanation: "Vehicle is moving on ego's lane, so ego moves" $S = 0$
Counterfactual Explanation: "If ego must stop, the following should be happening: a vehicle stopped on ego's lane" $S = 0$
c.) Right lane change action explanation:
Factual Explanation: "Traffic light is green on ego's lane, so ego moves straight" $S = 0.83$
Counterfactual Explanation: "If ego must move straight, the following should be happening: a vehicle stopped on outgoing lane" $S = 0.83$
d.) Left lane change action explanation:
Factual Explanation: "Ego's next goal is to move to the right lane, so ego moves to the left lane" $S = 0$
Counterfactual Explanation: "If ego must move to the right lane, the following should be happening: a vehicle is braking on ego's lane" $S = 0.65$

Figure 5.6: Sample factual and counterfactual explanations for the four actions along with entropy scores: (a) Ego stops due to traffic light; (b) Ego moves as the vehicle in front moves; (c) Right lane change action is misclassified as a move action, and thus its explanation would better be suited for a move action; (d) Explanation is based on plan. Ego moves to the left lane and then plans to counteract this by moving to the right lane. Explanations with lower entropy seem to be slightly more plausible.

Scenario A In the stop scene depicted in Figure 5.6a, the ego vehicle was stopping in front of a vehicle in a stop state due to the red traffic light. The explainer selected only the traffic light as the course for the ego vehicle's stop action; therefore fulfilling the selective requirement. The factual explanation follows the sequence uncovered in Section 5.3.3, i.e., *announce observation* → *announce plan/action*. A counterfactual explanation is also generated. In this case, the desired counterfactual action is *move* based on the plan of the ego vehicle. In this example, we placed a constraint on the EgoPlan feature so that the ego's plan is not modified when generating counterfactuals.

Scenario B The scene in Figure 5.6b depicts a move action. The ego vehicle keeps moving as long as the vehicle ahead moves. Its future plan is a stop; according to the counterfactual explanation, this would happen if a vehicle stops in front of the ego in the ego's lane.

Scenario C In the right lane change scene depicted in Figure 5.6c, a right lane change action was misclassified as a move action. The entropy value is highest

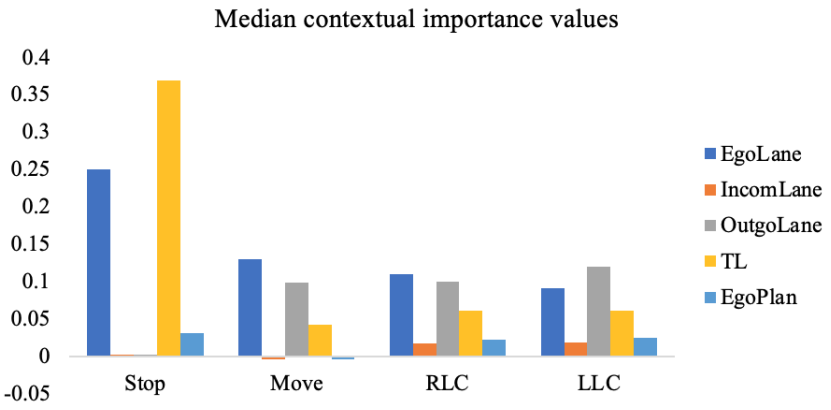


Figure 5.7: Overall, traffic lights, observations on the vehicle lane, and observations on the outgoing lane have the highest contribution to the actions of the ego vehicle.

in this example compared to the other examples. Hence, the explanation might not be as accurate as others.

Scenario D The scene in Figure 5.6d depicts a left lane change action. The ego provided a factual explanation based on its next plan. While the counterfactual explanation might be plausible, the current ego lane runs into roadside buildings so ego had to change lane to the right as soon as possible. This limitation occurred as we did not consider off-road objects and static objects in our study.

Finally, we computed the median contextual importance (CI) scores for the correct prediction from all the test sets for the four actions; see Figure 5.7. Overall, traffic lights, observations on the vehicle lane, and observations on the outgoing lane have the highest contribution to the actions of the ego vehicle.

Quantitative Results

We measured the degree of similarity between the generated factual explanations with ground truth explanations using the BiLingual Evaluation Understudy (specifically cumulative weighted BLEU-4) and The Recall-Oriented Understudy for Gisting Evaluation (specifically the weighted LCS ROUGE-W). The mean similarity score was calculated based on the average median entropy (0.95) for each class. Low entropy means that the model has high confidence and vice-versa. Figure 5.8 shows the distribution of the entropy values per class. The model has the highest

confidence for stop action predictions. From Table 5.6), explanations for the stop

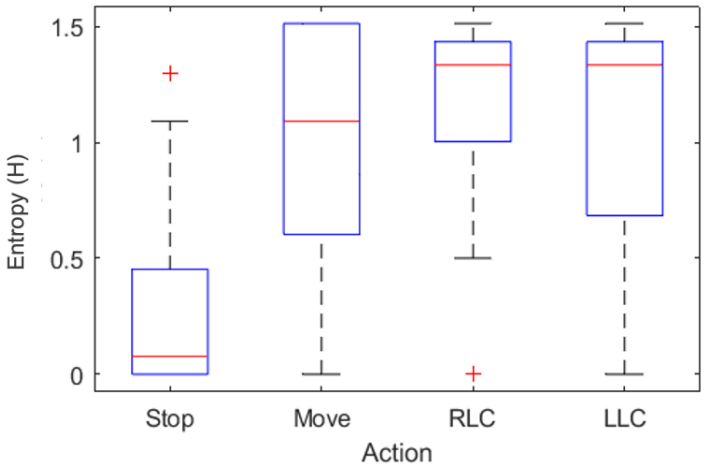


Figure 5.8: Certainty of the explainer per class with overall minimum = 0, overall maximum = 1.46, overall median = 0.95.

and move actions were generally more similar to the ground truth than lane change actions. Explanations generated for high confidence predictions ($H \leq .95$) are better for stop and right lane change (RLC) actions.

	BLEU-4		ROUGE-W	
	$H \leq .95$	$H > .95$	$H \leq .95$	$H > .95$
Stop	.648	.561	.732	.653
Move	.537	.693	.763	.783
RLC	.594	.348	.697	.462
LLC	.498	.568	.627	.672

Table 5.6: Comparing generated factual explanations with ground truth explanations. We choose a median entropy value of .95. Similarity scores seem to increase slightly with lower entropy values. $Min(H) = 0, Max(H) = 1.46, Median(H) = 0.95$.

Qualitative Results

We randomly selected 12 four-second videos. We ensured that each ego action had 3 examples. We presented these videos with their corresponding generated explanations to 20 human judges (10 males and 10 females) all with driving licences and driving experiences in the UK. These judges were recruited through the Prolific platform. We asked the participants to rate both the factual and counterfactual

explanation on the scale 0...3 (3: correct, 2: minor error, 1: major error, 0: completely wrong). Factual and counterfactual explanations for stop actions had the highest ratings for correct explanations (mean frequency of 14 and 13, respectively). The lowest rated factual and counterfactual explanations were those for right lane change (mean frequency of 2.333) and move actions (mean frequency of 6.667), respectively (See Figure 5.9).

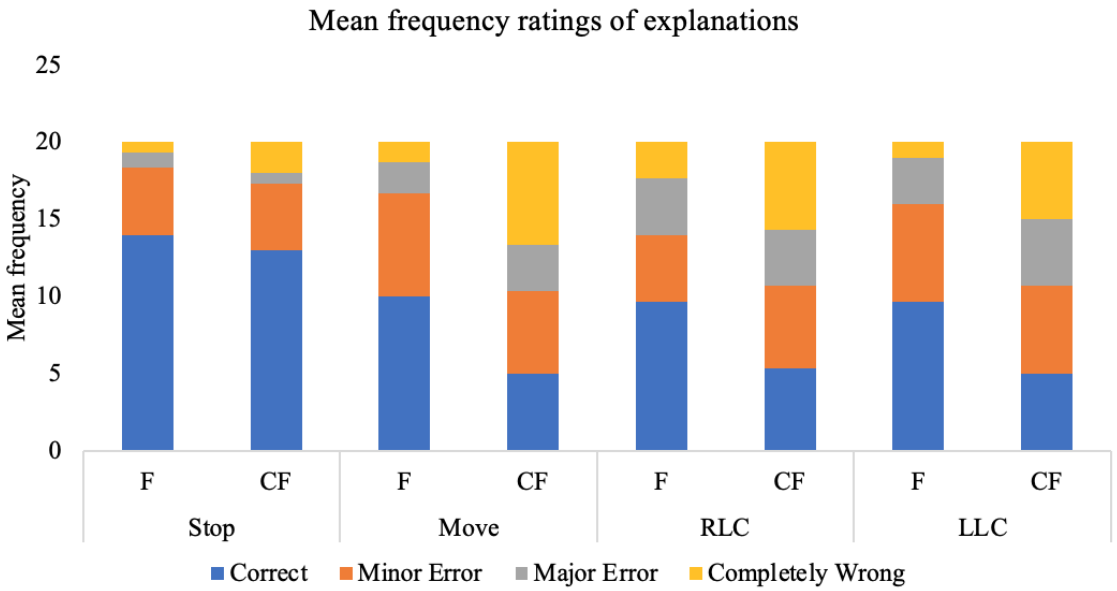


Figure 5.9: Participants' ratings for the factual and counterfactual explanations for each class.

5.4 Discussion

This research, in general, shows the potential for the realisation of automated explainers that can provide automated driving commentary to assist passengers to make better judgements about AVs, and as well, increase the driving knowledge of learner drivers. The experiment in Case Study 2, in particular, shows how a tree-based model could be utilised for ego vehicle navigation action prediction, as well as used as a surrogate model to explain an ego vehicle's decisions. Explanations are generated following the tree-traversal and search explanation algorithms we have proposed.

There are limitations in the current work; the navigation model only predicted four high-level actions (stop, move, right lane change and left lane change). There were also performance limitations in the tree-based model (Figure 5.5) as evidenced in the issue pointed out from Scenario C and D in our experiment. This model performance limitation is due to the limited size of the collated dataset, especially for the lane change classes. Model performance could be improved with more examples of the rarer classes. Moreover, increasing the dimensions of the input space to also encode key static road elements might improve the model accuracy. While the model was trained on a dataset that was collected from an ego vehicle driven by a human, the explanation method and concepts in this chapter are transferable to AVs. This is so since, in practice, AVs are often times pre-trained or trained on data collected from human driving behaviours.

Apart from the insights discussed in Section 5.3.3, there are many more insights to learn from how humans explain events to benefit future research. For example, early on, during the ride in the field study, the driver provided basic information about his driving preference. He stated his lane preference and his use of the side mirrors. We suggest that the driving preferences programmed into AVs be provided to the passengers at the start of a ride. It is important for an AV to be able to detect special vehicles, e.g., foreign trucks (from other countries with different traffic rules), and anticipate their actions as the driver did throughout the driving exercise. Explanations should therefore be able to reference these subtle differences. As some of the driver's comments reflect that he made decisions based on other agents' distances, relative distance metrics between an AV and other road agents may be relevant for explanation purposes. Although challenging, we think that explainers should model how humans explain (De Graaf & Malle, 2017). While high-fidelity human-centred studies with stakeholders in the loop may be expensive and challenging to set up in autonomous driving research, they are beneficial for eliciting requirements and learning preferences to inform the development of more robust explainer systems.

5.5 Conclusion

In this chapter, we set up experiments to assess the algorithms defined in Chapter 4. These algorithms have increased transparency compared to other natural language explanation generation algorithms which utilise deep neural networks. They also indicate confidence scores for the predictions of the tree-based model being explained.

We applied FE1 (Algorithm 1) and CFE1 (Algorithm 2) on a collision risk prediction case study. Tree-based models were trained on the Lyft prediction dataset, and the models' predictions were explained using the algorithms. FE2 (Algorithm 3) and CFE2 (Algorithm 4), which are improvements on FE1 (Algorithm 1) and CFE1 (Algorithm 2) respectively, were applied on a navigation action prediction case. These algorithms were designed after an analysis of the driving commentary provided in the new dataset that we have provided (i.e., the SAX dataset). The improvements over the initial algorithms include the (i) ability to be selective in the presence of competing candidate explanations, (ii) the ability to generate counterfactual explanations while respecting user-defined constraints, (iii) and the provision of entropy scores to give a sense of the model's confidence. A tree model was trained on a new dataset that we have provided (SAX dataset), and posthoc explanations were provided for stop, move and lane change actions using FE2 (Algorithm 3) and CFE2 (Algorithm 4). These explanations are approximate explanations as the model from which it is generated from is only a proxy to the underlying logic of the ego vehicle. The SAX dataset with a new annotation scheme (e.g., driving commentary) is targeted towards building more explainable driving models. Results from our experiment indicate the possibility of explaining collision risks and navigation actions of intelligent vehicles in the real world. In the next chapter, we describe a highly AV immersive user study to assess the effects of the explanations generated (using our explainer algorithms) on AV passengers.

6

Effects of Explanation Specificity on AV Passengers

Contents

6.1 Introduction

In this chapter, we use a driving simulator, an automated auditory explainer adapted from the previous chapter, and a virtual reality headset to investigate the effects of transparency in highly automated vehicles with fallible perception systems.

When deployed in the real world, AVs may not always have perfect scene understanding—due to their imperfect perception systems—leading to an impact on generated explanations. Informing operators about the system's inherent fallibility could help them identify when they may be required to take over the system, thereby improving safety (A. Kunze et al., 2019a). Previous studies have investigated various means for conveying information, such as visual uncertainties to drivers using augmented realty (Colley et al., 2020; Colley, Eder, et al., 2021; Colley, Krauss, et al., 2021; Colley et al., 2022; A. Kunze et al., 2018), peripheral awareness display (A. Kunze et al., 2019b), and visual animation (A. Kunze et al., 2019a). While this transparency is critical for AV operators, developers, and incident investigators, it is unclear whether passengers would prefer such high transparency that includes subtle errors (e.g., identifying a van as a bus) that may be inconsequential to the planning and the overall action of the AV. Hence, the question as to the level of AV transparency required, especially for in-vehicle passengers needs to be addressed. Moreover, as passengers are expected to be able to engage in other activities during a ride, the visual mode of communicating awareness to passengers might be futile in conditions where their attention is required. Hence, auditory feedback and/or vibro-tactile feedback (A. Kunze et al., 2019b) are as well needed.

In this study, we used a driving simulator, an automated auditory explainer, and a virtual reality headset to investigate the effects of transparency in highly automated vehicles with fallible perception systems. We use the term 'low transparency' to mean the provision of abstract (i.e., vague) auditory explanations that conceal perception system errors, and 'high transparency' to mean the provision of specific (i.e., detailed) auditory explanations that expose all perception system errors.

This study contributes to the body of knowledge in explainable autonomous driving and human-machine interaction by providing:

1. an improvement on the interpretable explanation techniques proposed in Chapters 4 and 5 to fall-back to using a set of rules (based on simulation logic) where the data-driven model predicts the wrong actions. Auditory support is also added to read out textual explanations.

2. A new use case of varying levels of auditory natural language specificity in the presence of varying degrees of AV perception system errors.

3. Findings on whether high AV transparency,—through specific explanations—though critical to operators, is desired by passengers from an AV with a fallible perception system.

6.1.1 Research Questions

1. Given varying levels of perception system errors, how do natural language explanations influence passengers' perceived safety?

    - **H1.1 - Perceived Safety.** *Low transparency yields a higher perception of safety in an AV with perception system errors.* We hypothesise that passengers feel safer in a low transparency AV, even though it provides abstract explanations. People sometimes seek the truth, but most prefer views that agree with their expectations (Hart et al., 2009). Hence, specific explanations might expose perception system errors which might be against the views or expectations of the passengers. Moreover, placebo explanations have been shown to have similar positive effects as real explanations on people (Eiband et al., 2019).

    - **H1.2 - Feeling of Anxiety.** *Passengers' feeling of anxiety increases with increasing perception system errors in a highly transparent AV.* We assume that there is a link between perceived safety and the feeling of anxiety (Davidson et al., 2016; Quansah et al., 2022). Therefore, explanations frequently referencing misclassified actors would create an unsafe feeling which can result in an increased feeling of anxiety.

- **H1.3 - Takeover Feeling.** *Given highly transparent AVs, passengers' are more likely to develop the feeling to takeover navigation control from the AV which has higher errors in its perception system.* Although passengers were not meant to takeover control in this study, we expected that they would conceive the idea of taking over control from the AV when they repeatedly receive 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 (N. T. Chen et al., 2014; Hepsomali et al., 2017). Participants' fixation points and saccades should correlate with anxiety.

6.2 Passenger Study

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

6.2.1 Participants

We publicised the call for participation on various online platforms, such as the callforparticipants platform, university mailing groups, university Slack channels, research group website, and social media. 49 participants booked appointments to attend. Participants were informed about the possibility of getting motion sick or nausea. Five participants cancelled before the experiment, three missed their appointments, and we excluded records from two participants who did not complete

the experiment. 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. 32 (82.05%) participants had no prior experiences with autonomous systems, 4 (10.26%) had experiences with semi-automatic driving systems, such as lane assists, and 1 (2.56%) had experiences with automated driving systems, such as highway and take-over driving assistants, and 2 (5.13%) had experiences with autonomous vehicles in research contexts. 6 (15.38%) of the participants had used a virtual reality headset for a driving game or driving experiment in the past, while 33 (84.62%) never had such an experience.

6.2.2 Apparatus

Figure 6.1: Driving simulation setup for the study. Setup included a: VR headset, steering wheel, brake and acceleration pedals, screen, and arcade seat.

Hardware

The hardware setup is shown in Figure 6.1. We conducted the experiment in a non-moving 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 monitor 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.

Driving Software

Software architecture is illustrated in Figure 6.2. 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 autonomous driving, 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. The software was powered by a PC running Windows 10 operating system with high-capacity CPUs and a GPU.

We created 3 different driving scenarios within a city (Town10HD, a high-definition urban city in Carla) using DReyeVR software, where each scenario is a drive along a different predefined path. Scenarios were about 4 minutes long with 55 vehicles (including cyclists, motorbikes, cars, vans, trucks, and emergency vehicles) and 20 pedestrians. All scenarios comprised different scenes (as listed in Table 6.1) with most similar to the explanation critical scenes described in (Wiegand et al., 2020).

Explainer Software

As shown in Figure 6.2, we adapted the explainer system that we proposed in Chapter 4. As the algorithms are data driven, we incorporated a rule-based logic that acts as a fallback when the data-driven method fails or makes an incorrect prediction. We know when a prediction is incorrect as we have ground truth observations from the simulator. We used this explainer system to generate preliminary explanations for the created scenarios. The explainer processes ground truth detections from either Carla or DReyeVR, predicts AV's action and generates a corresponding explanation for the prediction. While Wintersberger et al. (2020) suggested the

Table 6.1: Description of events and corresponding explanations provided by our explainer. Observations and causal explanations are announced to passengers' hearing. Texts in square brackets are placeholders for the information processed by the perception system. *actor type* is the type of actor e.g., car, traffic light, etc. *actor type1* and *actor type2* are used to differentiate between two different actors when they appear in one explanation. *colour* is the colour of the actor where it is necessary, e.g., *green* or *red* for a traffic light.

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.	[actor type] ahead on my lane.	Stopping because [actor type] 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.	[actor type] crossing my lane.	Stopping because [actor type] 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.	[actor type] ahead on my lane.	Changing lane to the [right/left] because [actor type] stopped on my lane.
LaneChangePlan	AV follows a leading actor, at some point AV indicates accordingly and changes to the lane on its plan.	Changing lane to the [right/left].	None
SignalisedJuncTurn	AV is turning right while indicating accordingly at a signalised intersection and turns into the same direction as another actor, crossing straight initially from a lateral direction.	None	None
StopSignalNoActor	No actor ahead of the AV at a signalised intersection with a red traffic signal. AV decelerates and stops.	[colour + actor type] ahead on my lane.	Stopping because [actor type] is [colour] on my lane.
StopSignalWithActor	AV stops behind an actor at a signalised junction or intersection.	[actor type1] ahead on my lane. [colour + actor type2] ahead on my lane.	Stopping because [actor type1] stopped on my lane; [actor type2] is [colour] 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 [actor type] is [colour] on my lane.
MovSignalWithActor	AV starts moving from a stop state behind a moving actor at a signalised junction or intersection.	None	Moving because [actor type] is [colour] on my lane.

types of traffic elements to be included in visual explanations based on a study on user preferences, our proposed explainer picks up traffic elements that the driving

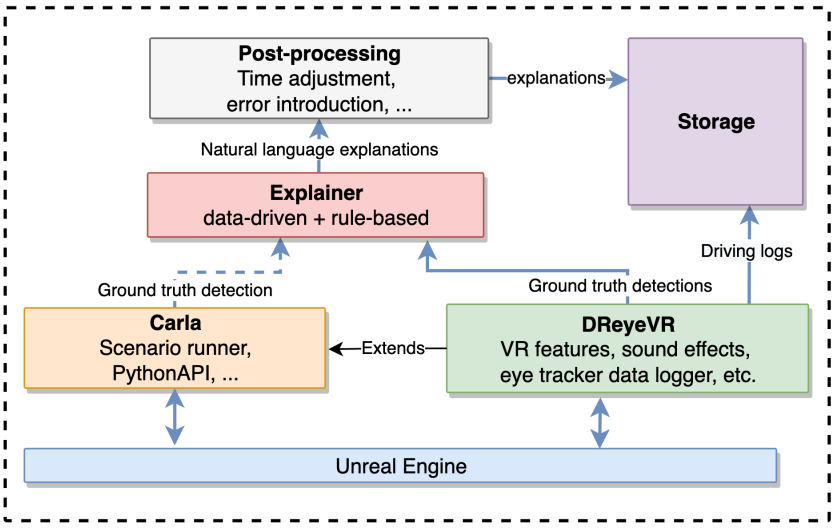


Figure 6.2: High-level architecture of our simulation system. DReyeVR uses Unreal engine and extends Carla simulator which also builds on Unreal engine. DReyeVR extends Carla by adding VR functionalities, vehicular and ambience sounds, eye tracker data logging, and additional sensors, among others. Our explainer model which is both rule-based and data-driven can receive ground truth data from either Carla or DReyeVR and generate explanations for predicted actions. The post-processing script allows us to modify the generated explanations as we desire.

model deemed important for its driving decisions.

We performed post-processing operations on the generated explanations. Post-processing operations included (i) fine-tuning some of the explanations (ii) modifying explanations' timestamps to make them come at the right time (iii) interchanging the actors that are referenced in the explanations to reflect various degrees of perception system errors.

6.2.3 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. The aim of the drive was only to familiarise participants with the driving simulation environment and to satisfy their desire to experience manual driving in a simulation environment. A within-subject design was done as our sample size was not large enough for a between-subject study. Moreover, we

wanted to avoid any potential co-founding factor of between-individual differences in a between-subject design.

Independent Variable

Combinations of transparency level (low and high) and AV perception errors (low and high) were done to obtain the independent variable *Scenarios*. The first scenario (*Abstract* scenario) comprises abstract explanations indicating low transparency and an undefined amount of perception system errors. The second scenario (*Specific(5)* scenario) comprises specific explanations indicating high transparency and 5% amount of perception system errors indicating low error degree. The third scenario (*Specific(50)* scenario) comprises specific explanations indicating high transparency and 50% amount of perception system errors indicating high error degree. Scenarios

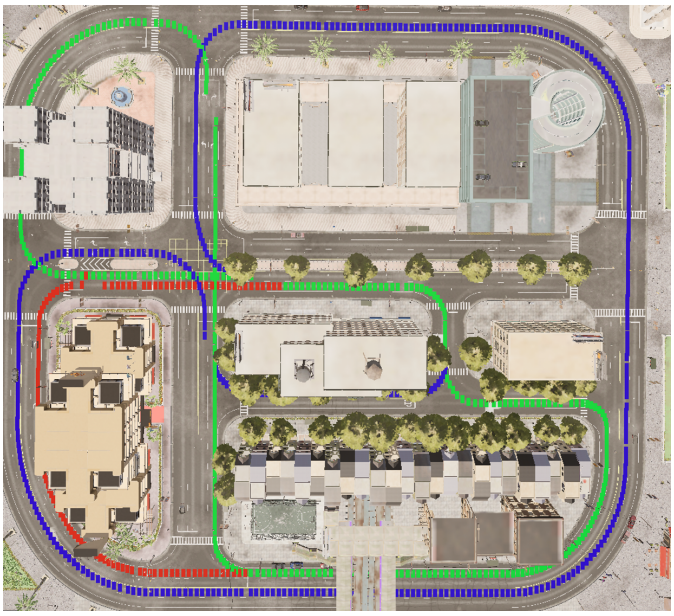


Figure 6.3: Scenario routes. **Red**: Abstract, **Green**: Specific(5), **Blue**: Specific(50). Each route is a loop and overlaps with others at some points.

were carefully designed to include different driving conditions that are obtainable in the real world (See Table 6.1). The scenario routes are shown in Figure 6.3.

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 6.1. 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 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.

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 6.1. 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 and 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 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 6.1. 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.

Note that all three scenarios were designed so that the AV perception 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 perceptual errors in AVs. Hence, it was necessary to introduce artificial errors of varying degrees (low and high). The non-influence of AV perception 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 6.1. In all the scenarios, the AV maintained a speed below $30mph$, the recommended speed limit in urban areas in the UK. See Figure 6.4 for sample scenes from each scenario and their corresponding explanations.

Dependent Variables

There were six dependent variables: *Perceived Safety, Feeling of Anxiety, Takeover Feeling, Fixation Divergence, Saccade Difference,* and *Button Presses.* These

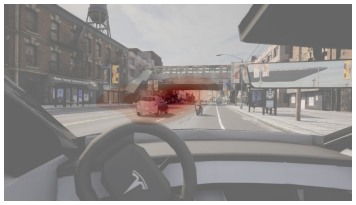
variables were categorised into two (psychological factors and behavioural cues) for easy analysis and reporting.

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).

(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.'**

Figure 6.4: 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 (Figure 6.4a), all movable/dynamic non-human actors are referred to as 'Vehicle'. Thus, a cyclist was referred to as a vehicle. Figure 6.4b depict 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 (Figure 6.4c), the AV's perception system misclassified a pedestrian as a cyclist. The fine-grained/specific explanation provided exposed this error.

Behavioural Cues We also used *Button Presses, Fixation Divergence*, and *Saccade Difference* as additional metrics. *Button Presses* were used to express 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.

6.2.4 Procedure

The experiment's procedure is illustrated in Figure 6.5. The researcher sent an information sheet to the participants before they arrived. The researcher welcomed the participants and stated the aim of the experiment. The researcher asked the

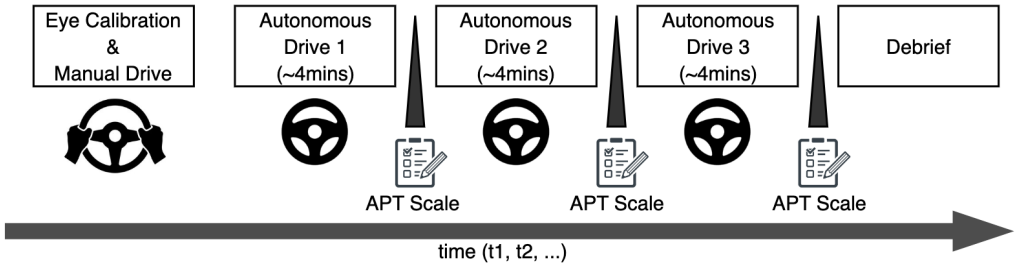


Figure 6.5: Study procedure. Eye calibration was done with the VR headset; participants drove for two minutes, participants experienced each of the 4 mins scenarios in counterbalanced order and completed the Feeling of Anxiety, Perceived Safety, and Takeover Feeling Questionnaire (APT Scale) in between each scenario. Participants were debriefed.

participants to complete a COVID screening form, an optional photography consent form, and a questionnaire regarding demographics.

The researcher then introduced the driving rig and explained the next steps, which involved eye calibration with the VR headset and a manual driving session which lasted for 2 minutes. The researcher stated the aim of the manual driving exercise: '[...] the next step is for you to drive in the simulator. The aim is to get you familiar with the simulator and the driving environment. It is not part of the main experiment.[...]'

When the manual driving exercise was completed, the researcher took the VR headset off the participant and explained the aim and the procedure of the main experiment. The instructions from the researcher 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 put back the VR headset on the participants and launched the scenarios. Complete counterbalancing was done for the Scenario treatments. Specifically, we had six different orders of scenarios upon permutation.

Each participant experienced the scenarios in one of the six orders. Approximately six participants experienced the scenarios in the same order.

We encouraged participants to rest for a while after each driving experience with the VR headset off their heads. There was a short debrief session after the study, after which the participants were handed a £10 Amazon gift card. The experiment lasted for approximately 50 minutes. The researcher participated in the experiment and experienced all three scenarios. The researcher always focused on the lane ahead and fixated on the actors that the explanations were referencing. Neither the erroneous nor abstract explanations influenced the researchers' focus, as the researcher always focused on the lane and the actors/obstacles that influenced the AV's actions, even when the explanations said otherwise. This was possible because the researcher was already very familiar with all the scenarios. The data from the researcher was used as a reference/ground truth. Note that the researcher whose data was used as ground truth moved between fixation points at about normal human saccadic velocity. Normal saccadic velocity reaches 300—400°/seconds (Raab, 1985; Wilson et al., 1993).

6.3 Quantitative Results

6.3.1 Psychological Factors Analysis

To test our hypotheses listed in Section 6.1.1, we analysed the data from the three APT questionnaires. A latent variable (*Feeling of Anxiety*) was formed from the means of the responses from AVAM Items 19–21. Another latent variable (*Perceived Safety*) was formed from the means of AVAM Items 24—26. We calculated the Cronbach Alpha (α) for the independent variables from which the latent dependent variables were formed to see if they had adequate internal consistency. Results with adjusted p-value less than 0.05 ($p < .05$) are reported as significant. p-values were adjusted using Bonferroni corrections, where the calculated p-values were multiplied by the number of scenarios. These corrections were made to reduce the chance of false positive errors (Type 1 errors). Kolmogorov-Smirnov, Shapiro-Wilk, and Anderson-Darling tests indicated a normality violation in the *Feeling of Anxiety*,

Table 6.2: 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
Vague	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**

Perceived Safety and *Takeover Feeling* factors. Therefore, a Friedman test was performed for these dependent variables. See Table 6.2 and Figure 6.6.

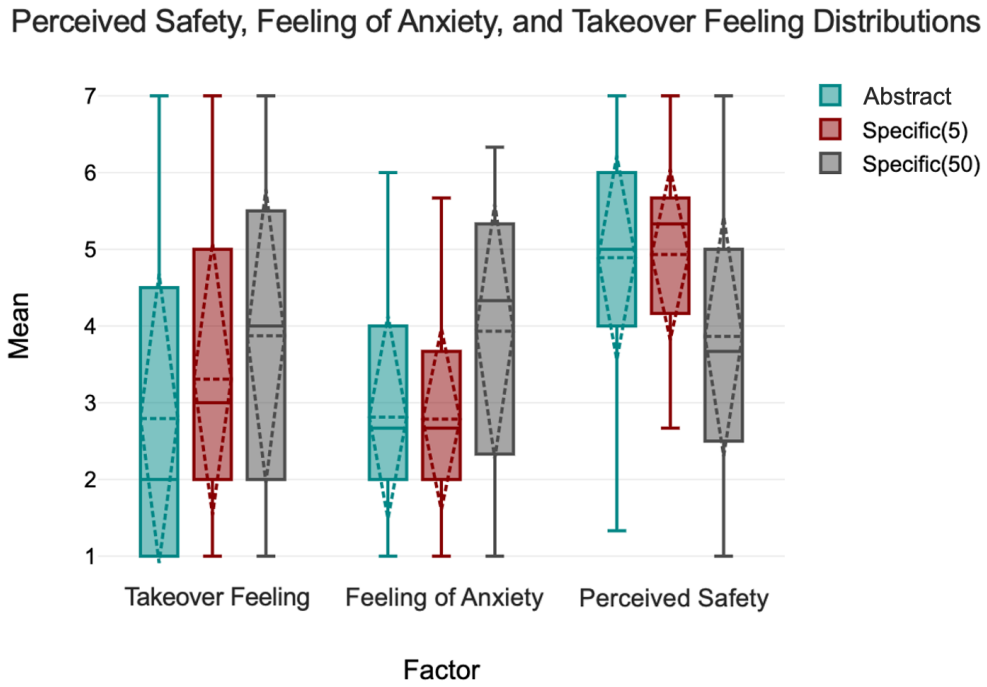


Figure 6.6: Perceived safety, feeling of anxiety, and takeover feeling distribution. Perceived safety is highest in the Specific(5) scenario, the feeling of anxiety is highest in the Specific(50), and takeover feeling is lowest in the Abstract scenario.

H1.1 - Perceived Safety

Low transparency yields a higher perception of safety in an AV with perception system 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 6.2. Therefore, there was no sufficient evidence in support of hypothesis H1.1.

H1.2 - Feeling of Anxiety

Passengers' feeling of anxiety increases with increasing perception system 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 6.2). Hence, there is strong evidence in support of hypothesis H1.2.

H1.3 - Takeover Feeling

Given highly transparent AVs, passengers' are more likely to develop the feeling to takeover navigation control from the AV, which has higher errors in its perception system. 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 6.2).

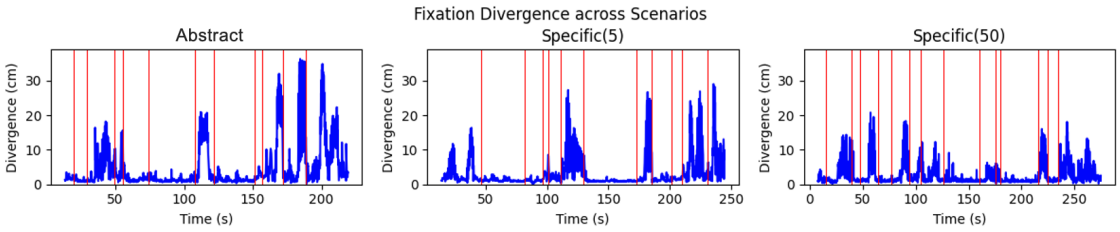


Figure 6.7: Fixation divergence across scenarios. While *Specific(5)* had the highest mean fixation divergence, *Specific(50)* had more frequent high fixation divergences. Red vertical bars represent the positions in time where causal explanations were provided.

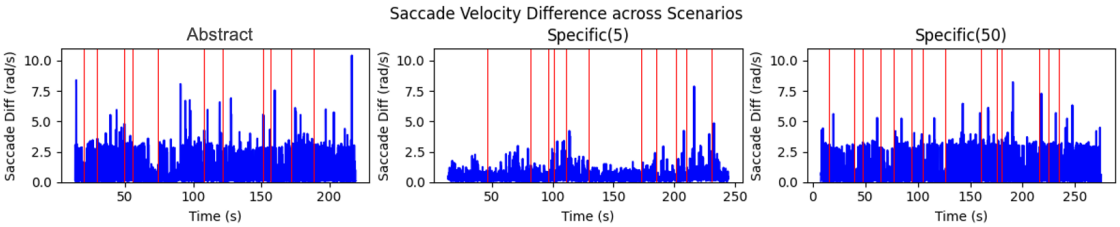


Figure 6.8: Saccade velocity difference across scenarios. *Specific(5)* had the lowest mean saccade velocity difference while the *Abstract* scenario had the highest. Red vertical bars represent the positions in time where causal explanations were provided.

6.3.2 Behavioural Cues Analysis

H2.1 - Visual Responses

Visual feedback from passengers correlates with passengers' anxiety. The ground truth data from the researcher was used at this point. Euclidean distances between participants' fixation points and the ground truth fixation points were estimated 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 Figure 6.7. 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 Figure 6.8.

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 6.3 for descriptive statistics. Also, there was a significant difference between *Abstract* and *Specific(5)* with respect to *Saccade Difference* (adjusted p-value of $< .001$). See Figure 6.4 for sample scenes from

Table 6.3: Descriptive statistics from the haptic and Visual responses.

	ButtonPress $H(2) = 15.44, \boldsymbol{p < .001}$			**Fixation Divergence** $H(2) = 20.67, \boldsymbol{p < .001}$			**Saccade Difference** $H(2) = 15.35, \boldsymbol{p < .001}$		
	Mean	SD	Mean Rank	Mean	SD	Mean Rank	Mean	SD	Mean Rank
Vague	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

each scenario with the generated explanations. All the participants' gaze points are plotted as heatmaps over the screenshots.

Haptic Response

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 statistical 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 Figure 6.9 for behavioural cues results.

6.4 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

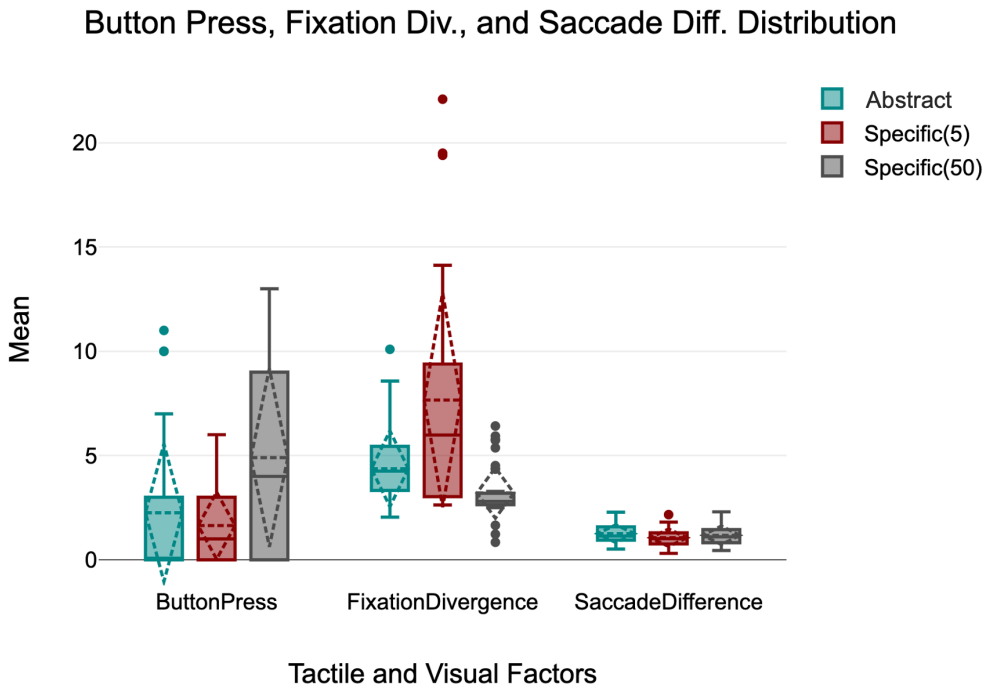


Figure 6.9: Button presses, fixation divergence and saccade difference distribution. Button presses are fewer in the *Specific(5)* scenario. Fixation divergence is highest in the *Specific(5)* scenario, and saccade difference is lowest in the *Specific(5)*.

they received during the ride. Table 6.4 and Figure 6.10 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

Table 6.4: 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
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
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

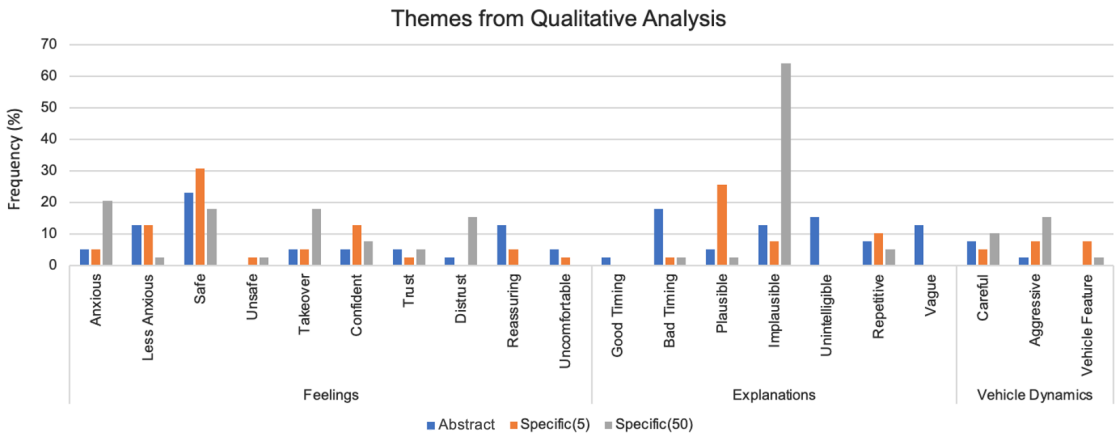


Figure 6.10: Themes derived from the thematic analysis of the qualitative data from participants. Frequency is expressed in percentage of the total number of responses in each scenario.

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.

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 next directions. '*The indicator sound was nice to hear. [...]*'—CAND6 (*Specific(50)* scenario).

6.5 General Discussion

This work investigated the effects of two levels of explanation specificity (abstract and specific) in the presence of two different degrees of perceptual 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. The experiment was conducted using an immersive autonomous driving simulator and a VR headset. Our results corroborate prior studies by showing that intelligible explanations create positive experiences on users in autonomous driving (Ha et al., 2020; M. Faas et al., 2021; Omeiza, Web, et al., 2021; Schneider et al., 2021), but only when the AV's perception system errors are low in our case.

Psychological Effects

Hypothesis 1.1 - Low transparency yields higher perception of safety Against expectations, participants showed a higher feeling of safety for the *Specific(5)* scenario. This signals the preference for specific explanations in an AV with significantly minimal perception system errors. On the other hand, too detailed explanations could be thought to be verbose and repetitive. A couple of participants thought this of the *Specific(5)* scenario. Thus, a good balance between the specificity of explanations (or transparency generally) and the cognitive load on passengers is essential (Poursabzi-Sangdeh et al., 2021).

As observed from Figure 6.6, highly transparent AVs with a high degree of perception system errors evoked lesser feelings of safety in passengers. However, a few participants—from the qualitative response—had positive feelings even in the presence of these errors. They applauded the vehicle for detecting obstacles and responding appropriately to them. For this category of people, the type of obstacle does not really matter as far as the AV makes the right decision.

Hypothesis 1.2 - Feeling of Anxiety increases with increasing perception system errors. Drivers' anxiety has been shown to increase when they use AVs (Koo et al., 2016). In our study, we expected the passengers' feelings of anxiety to increase with the increase of perception system errors in an AV. This was the case as there was a significant difference in the feeling of anxiety between the *Specific(5)* and *Specific(50)* scenarios. Since participants' perceived safety was highest in the *Specific(5)* scenario (based on the result from Hypothesis 1.1), the feeling of anxiety should be lowest in the *Specific(5)* scenario as we assume that perceived safety and anxiety feelings are related based on the finding from Dillen et al. (2020). While Dillen et al. (2020) mainly focused on how AV driving style influences passengers' anxiety and comfort, they noticed that the feeling of anxiety for some of the participants was influenced by some in-vehicle features, such as the rotating steering wheel. This was reflected in the comments from participants (CAND21).

Hypothesis 1.3 - Takeover feeling increases with the increase in perception system error. While there was a significant difference between takeover feeling in the *Abstract* and *Specific(50)* scenarios, there was no significant difference between scenarios *Specific(5)* and *Specific(50)* where perception system errors were exposed. Hence, this hypothesis was rejected. So it could be inferred that the feeling of anxiety due to increased AV perception errors does not necessarily evoke passengers' urge to takeover control. This contrasts the suggestion by (Terken & Pfleging, 2020) that full human-out-of-the-loop automated driving may not be welcomed by the users of the technology, and hence, argued for a shared control between the vehicle and the user. It is worth noting that the authors' conclusions made in (Terken & Pfleging, 2020) were based on conceptual analysis without empirical support.

Behavioural Cues

Hypothesis 2.1 - Visual signal correlates with anxiety. While Hepsomali et al. (2017) draws a connection between anxiety and distraction, we found no correlation between fixation points divergences and the feeling of anxiety across scenarios. This might be because people most times do have different priorities

in terms of where to fixate. What might be appealing to CAND10 might not be appealing to CAND11. Some participants might have paid more attention to the city fabric than the area of focus hinted at by the explanations, thereby, reducing their situational awareness. As shown in Figure 6.7, there were more frequent high divergences in fixation points between the participants and the ground truth reference in scenario *Specific(50)* as compared to *Specific(5)* and *Abstract* scenarios. Participants' attention might have been wrongly directed to the wrong actors/obstacles due to the actor misclassifications evidenced in the explanations. However, we had a couple of extremely high divergences in the *Specific(5)* scenario. While participants might have maintained normal focus most of the time, interesting sites/actors might have caught their attention intermittently. The *Abstract* scenario created nearly similar fixation effects as the *Specific(5)* scenario. This indicates that the explanations might have been more helpful in the *Specific(5)* and *Abstract* scenarios than in the Scenario(50) scenario.

While Dillen et al. (2020) drew a connection between eye movement entropy and anxiety, no significant correlation between saccade difference and the feeling of anxiety was observed in our data. Saccade velocity difference was lowest in the *Specific(5)* scenario. This might be a sign of less distraction and/or confusion as saccade velocity indicates how fast people move between fixation points. Saccade velocity difference was highest in the *Abstract* scenario. Perhaps this could be a search indication as explanations were not specific enough to quickly direct participants' gaze.

6.5.1 Practical Implications

While we assumed from the outset of this chapter that passengers may not want specific explanations that provide error details, the study suggested otherwise. Passengers prefer specific explanations from an AV with a near perfect perception system. Since high perception system errors negatively impact anxiety based on our study, manufacturers and regulators should ensure the design of highly

transparent AVs with very high perception and decision accuracy so as to reduce the effect of errors on the passengers.

Though the misclassification errors did not have consequences on the AV's action in our study, obstacle type is critical in determining the magnitude of inputs to be sent to the AV actuators as dynamic obstacles have varying manoeuvre capabilities. While the AV might not have made any glaring incorrect navigation decision due to the low consequences of the perception system errors, in an extremely complex scenario, it needs highly accurate estimations of its environment to determine how much change in speed (acceleration) is required to overtake when the obstacle is a bicycle and to re-adjust when the obstacle is a motorcycle. Hence, transparency and accuracy should be considered hand-in-hand.

While visual feedback from experiments of this type can be useful for inferring psychological and/or behavioural effects on people, we strongly advise that they be complemented with other measurements, such as surveys to reach more confident conclusions.

6.5.2 Summary of Findings

The following findings were made:

1. Passengers felt safer in highly transparent AVs with minimal perception system error compared to a low transparency AV which provided abstract explanations to conceal perception system errors.

2. Passengers' feeling of anxiety increased in the presence of high perception system errors in an AV.

3. While the feeling to takeover driving operation from the AV increased in the presence of perception system errors, there was no significant difference in the feeling to takeover navigation control across the different magnitude of these errors.

4. Finally, while visual patterns varied significantly across driving scenarios, no correlation was found between the feeling of anxiety and visual signals.

6.6 Conclusion

In this chapter, we conducted a within-subject lab study ($N = 39$) using an immersive driving simulator to investigate: (1) passengers' perceived safety, feeling of anxiety, and takeover feeling in AVs based on explanations of different specificity from an AV with varying perception system errors; and (2) relationship between passengers' behavioural cues and their feelings during an autonomous drive. Our results showed that passengers felt safer under specific explanations provided by the AV with low perception system errors, even though abstract explanations concealed AV perception errors. The feeling of anxiety increased in the presence of perception system errors exposed through the provision of specific intelligible explanations. In addition, no correlation was found between behavioural cue from visual signals and the feeling of anxiety. Participants' control over the explainer and the AV's driving style was very minimal in this work. In future work, we would investigate the implications of providing passengers with the choice to personalise the explainer and the driving style of an AV during an autonomous drive.

7

Conclusions, Recommendations and Outlook

Contents

This thesis opened by drawing attention to the overarching need for intelligible, human-centred explanations in autonomous vehicles. Intelligibility and causal attributions are argued to be important properties of sound explanations in autonomous driving; exploratory studies were conducted to support this argument. Transparent methods for generating intelligible explanations with causal attributions in autonomous driving were proposed. These methods were deployed in simulation environments to examine the effects of the explanations they generate on humans. In this chapter, we first summarise key findings from the studies. We then discuss

the limitations of this work and reflect on the challenges associated with the wider application of this work. Finally, we set out a long-term vision to realise highly effective explainability in autonomous driving.

7.1 Summary of Results

As outlined in Chapter 2, there exists limited research on human-centric explainability in autonomous driving. Thus, we set out to address the gap by first investigating the different explanations and driving scenarios envisaged in highly automated driving environments. Our literature research in Chapter 2 helped us identify concepts and types of explanations that have been defined in social science, e.g., contrastiveness, social, counterfactual, causal attributions, among others. Moreover, different driving actions—e.g., goal-driven and stimulus-oriented actions—surfaced from the literature survey. Further, we identified the complexity challenge in the existing explanation methods that have been proposed for autonomous driving. Lastly, we observed a relative absence of comprehensive and specific regulations aimed at explainability in autonomous vehicles.

Chapter 3 expands on the identified explanation and driving scenario types and provides empirical evidence for the varying utility of the different explanations. A mixed-method study was conducted to explore the utility of the different explanations in the different driving scenarios. While using measures, such as intelligibility, accountability, and trust to assess the different explanation types in different scenarios, we discovered that explanations with causal attributions (compared to non-causal explanations) create a better understanding of driving actions, accountability and trust. These causal explanations are those triggered by a *Why, Why Not,* and *What If* investigatory queries. Recall that explanations with causal explanations are explanations that provide reasons for their current state based on external factors. In contrast, those without causal attributions (which we term 'non-causal explanations') only provide information about their state/action without supporting reasons. Hence, causal explanations are best suited for challenging scenarios, such as those that involve emergency vehicles

(emergency scenarios), collision (collision scenarios), and near-misses (near-miss scenarios). Interestingly, our results showed that people's perception of trust in autonomous vehicles declined after participating in the study. We attribute this to the complexity of the scenarios presented in the study, where a few collision cases were shown to the participants. This decrease in trust is evidenced despite the fact that the AV was not responsible for any of the collisions. When asked what they would desire in an explanation, participants generally desire intelligible, concise, and visually appealing explanations.

Chapter 4 makes an argument for more transparent and modular designs for AV systems and provides a conceptual design for an explainable AV, and as well as data structures and algorithms for generating intelligible explanations for AV. It starts by highlighting key requirements for AV architectures, arguing that an AV architecture should be such that it is not overly complex so that its high-level workings are expressible in clear natural language. It should be transparent, facilitate easy system auditing, and enable an easy incident investigation. On top of this, a conceptual modular framework for explainable AV was proposed with independent perception, planning, control and system management components. A tree structure was proposed to represent observations, road rules and actions of an AV to facilitate the generation of explanations. This tree structure provides a transparent way to represent driving scenarios over time. Two practical and important problems in autonomous driving were introduced; explainable collision risk prediction and explainable navigation decision prediction. Carefully designed transparent solutions were defined, and transparent algorithms were also proposed to generate intelligible natural language explanations for the predictions.

In Chapter 5, experiments were conducted to demonstrate the algorithms proposed in Chapter 4. The major contributions of this chapter are the introduction of a new dataset—the SAX dataset—for explainable autonomous driving research, and the successful generation and evaluations of explanations for collision risk predictions and navigation decision predictions. The SAX dataset was obtained from 9.5 hours of driving and is unique compared to the existing datasets. It differs in

that it provides very rich semantic information that is useful for designing explainable scene understanding models and explainable navigation decision prediction models. This dataset was used to train a tree-based navigation action prediction model whose predictions were explained by the explainer algorithms in Chapter 4. The generated explanations proved intelligible and mostly plausible based on results from quantitative and qualitative evaluations.

Further, Chapter 6 argued that, in reality, perception systems in AVs are not perfect. It, therefore, raised an interesting question as to what the effects of the exposure of perception errors, through explanations, are on passengers . Hence, the effects of explanation specificity (abstract and specific) in the presence of two different degrees of AV perception system errors (low and high) were investigated using a state-of-the-art virtual reality headset and a physical driving simulator. Visual (through eye tracking) and haptic (through button clicks) feedback were gathered as participants in the experiment responded to the actions taken by the autonomous vehicle and the natural language explanations provided. Through quantitative and qualitative analysis of the provided questionnaire data, and the visual and haptic data, various findings were made: First, it was discovered that specific explanations provided better positive effects on passengers compared to abstract ones. This was so despite the fact that abstract explanations concealed perception system errors. Second, passengers' perception of safety and anxiety levels are adversely affected by high perception system errors in an AV. However, these errors did not have any significant influence on their feeling to takeover control from the AV. Lastly, while visual patterns varied significantly across the different driving scenarios, no correlation was found between the feeling of anxiety and the visual signals from the participants.

7.2 Reflection on the Research Questions

Research question R1: *What type of driving scenarios primarily demand explanations and what type of explanations are appropriate for these scenarios?*

This question was addressed in Chapter 3. A thorough exploratory study was conducted to investigate the impact of different explanations in different autonomous driving scenarios identified from the literature. The degree of this impact—measured through the intelligibility, accountability, and trust objectives—provided pointers to the explanation types that are most useful in autonomous driving scenarios. In addition, near-misses, emergencies, and collision scenarios stood out as scenarios where explanations could be very useful. Explanations with causal attributions as well proved to be more beneficial.

Research question R2: *How can intelligible explanations of these types be generated automatically for AV actions in the identified scenarios?*

This question was addressed in Chapter 4 and Chapter 5, where a conceptual framework for an explainable AV was provided. Tree-based representations and data structures were also provided to represent driving information for easy explanation generation. Algorithms were provided to generate posthoc explanations for autonomous driving actions. In Chapter 5 specifically, the Lyft-Level5 dataset and the SAX dataset were introduced upon which experimentations of the proposed algorithms were performed using a collision risk explanation case study and an AV navigation explanation case study respectively. These algorithms were able to generate intelligible natural language explanations for both AV experts and AV passengers.

Research Question R3: *How would passengers react to explainable but fallible autonomous driving systems?* This question was addressed in Chapter 6 where the effects of explanation specificity (abstract and specific) in the presence of two different degrees of AV perception system errors (low and high) were investigated using a state-of-the-art virtual reality headset and a physical driving simulator. Results disclosed that passengers have more feeling of anxious in the presence of high-degree AV perception system errors exposed through specific explanations. Passengers felt safer in specific explanations with lower perception errors compared to the abstract explanations. This makes specific explanations preferred over abstract explanations, especially in AVs with high perception accuracy.

7.3 Limitations

There are limitations associated with this work.

7.3.1 Human Factor Limitations

Not so Social: According to Miller (2019), explanations are social in that they 'are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs.' Explainers have been designed to be only unidirectional in our user studies. Hence, the explainee cannot respond when provided an explanation. However, our algorithms could easily be adapted to be conversational by incorporating investigatory query/prompt processing capability.

Absence of Explainer Personalisation: In our work, we did not provide the opportunity for explainees to define their preference in terms of the frequency of explanations, mode of explanation presentation (e.g., audio or text), and the levels of details the explanations should report. These are key factors for personalisation which would be critical for the human-machine interfaces in explainable AVs.

Limited Stakeholders Considered: While we have argued—at the beginning of this thesis—that comprehensive explanations could be used to investigate incidents in autonomous driving, we have only focused on how explanations would be useful to in-vehicle passengers. As AV decisions don't only affect passengers, other stakeholders, such as incident investigators, auditors, and even external traffic agents who have some level of influence in the wider adoption of AVs, are worth attention in future research.

7.3.2 Technical Limitations

Conceptual Design: The modular framework that we have proposed for explainable AVs in this work is conceptual. Hence, further work is required to evaluate its practicability.

Limitations of Tree Structures: Just as the blackbox models have their limitations, our tree-based algorithms are affected by the limitations inherent in tree-based models. For example, tree-based structures can be computationally expensive to traverse as they grow in size. However, pruning operations can prevent them from exploding. This can also reduce the over-fitting effect associated with tree-based models.

Independent Consideration of Tasks: We experimented with the proposed algorithms to explain predictions in a collision risk assessment task and a navigation decision task. While these tasks were presented independently in this work, in practice, the output from one is meant to be fed into the other. We envision a case where the explainers would be able to provide explanations for a combination of these tasks. Explanations for navigation decisions should reference the risks avoided for making a particular navigation decision. This case was not considered in our research.

Explanations Faithfulness are not Guaranteed: As with many posthoc explainers, we cannot assure 100% faithfulness of the explanations, as in principle, the proxy/surrogate model is what is being explained. Hence, we qualify the explanations generated in the AV navigation experiment as 'approximate', which are still useful in creating positive effects on passengers. Further, as the ego vehicle in Chapter 5 was driven by a human, one might argue that the explanations were rather for the human's decisions which in theory, is impossible to explain. In any case, many datasets used in training (or pre-training) AV models were collected by an ego vehicle driven by a human driver. So the methods proposed in this thesis are transferable to actual AVs.

7.3.3 Regulatory Limitations

The vagueness concern expressed about the existing regulations in Chapter 2 is still an open challenge. There is an absence of details on the nature and specificity of the explanations to provide the different AV stakeholders, as explanation requirements

differ across stakeholders. Moreover, specifics as per the explainability requirements for each component of the AV stack are missing. This makes the realisation of transparent and explainable AVs difficult. We provide recommendations—in the next section—for AV regulators to consider.

7.4 Outlook

Having discussed the limitations of our work, we describe the nature of future works.

7.4.1 Human Factors

First, the conversational style of explaining is an important topic to explore. This would involve forming a mental model of an explainee and adapting the explanations to the beliefs of the explainee. The effectiveness of such explanations could be assessed using metrics like time to reach closure in an explanation cycle, and the 'quality' of questions that the explainee asks in subsequent similar situations. This is a fruitful area for future work.

Second, the personalisation of the explainer system is important for realworld use. The implication of providing passengers with the choice to define their preferences is an interesting question to investigate. This includes their choice to adjust the frequency of explanations, the mode of explanation presentation (e.g., audio or text), feedback mode, and the levels of detail in which the explanations should report. This is particularly important for accessibility purposes.

Third, more inclusive research in explainable autonomous driving is needed to demonstrate value for different stakeholders. Future research involves investigating the utility of explanation logs to incident investigators.

7.4.2 Technical Factors

From the technical perspective, realising the future research goals mentioned above would require technical efforts. The explainer systems need to be improved to be more robust to allow for the explanation of more actions and for easy personalisation.

New paradigm for posthoc explainability, that can guarantee 100% faithfulness of explanations, needs to be explored.

7.4.3 Regulatory Factors

Explainable AVs would not be realised if there are no clear and favourable regulations to drive them. Hence, this is a very important area and the gap pointed out in Section 7.3.3 needs attention. We have provided some recommendations based on the experience acquired from this research.

Explanations can help in assessing and rationalising the actions of an AV (*outcome-based*), and in providing information on the governance of an AV across its design, deployment, and management (*processed-based*). This aligns with the ICO guidelines for general AI systems (Information Commissioner's Office, n.d.). We suggest that regulatory guidelines for AV explainability should be set in line with these two goals. We have carefully adapted the ICO's guidelines and the ITU's comments on the 'Consultation Paper 3 A regulatory framework for automated vehicles' to provide recommendations.

Outcome-based explainability

To explain the outcomes of an AV, one must consider explainability at the perception level (i.e., what the AV 'sees'), decision level (i.e., how the AV plans paths and motion) and action or control level (i.e., how the AV acts on its plan or how it executes its decisions). We use these high-level terminologies: perception, decision, and actions to describe how explainability regulations could be made specifically for AVs.

Perception Explanations for a perception system can come in two forms:

1. An explanation capable of explaining the algorithms or software processes used to transform sensor data into a digital representation of the real-world and justification for such algorithms. In line with the Molly Problem earlier mentioned in Chapter 1, it should be possible to obtain information on how

Molly was represented digitally (sensor types and data transformations), and information about the circumstance, e.g., her location, position (maybe coordinates), and the time the representation occurred.

2. An explanation that provides information on what this digital representation contains (e.g., pedestrian, vehicles, road fabric) and the state of these objects (e.g., crossing, heading north, static). For example, if Molly was detected, it should be possible to receive information to explain and justify the process (e.g., detection and tracking algorithm or software) by which Molly was detected and tracked and the detection and tracking confidence levels.

Suggestion: In explaining outputs from a perception system, we suggest that AVs should be able to provide real-time data access (both onboard or remotely) to their digital representation of the 3D world (including semantic information) and the algorithmic processes applied to interpret this data when requested by authorised entities.

Decision This involves the provision of insights into the planning operations (behaviour planning and path planning) of the AV. The decision-making steps involve planning paths and motion/behaviour based on observations (through perception) from the environment and its structured knowledge about the environment. This planning requires reasoning and decision-making under several constraints, e.g., uncertainty about the environment's current and future state. It also involves identifying potential risks, evaluating them and finding measures to mitigate them.

These uncertainties are associated with the confidence level in the AV's predictions. AVs use prediction algorithms to predict their trajectory and that of other road participants and the risk of collision associated with different plans. In relation to the Molly Problem, to ascertain whether the AV had a good awareness of the environment upon which it made a decision, we must consider whether it was aware of the confidence levels of the models that detected and tracked Molly. It should be able to provide information on how confident it was about the subsequent steps or actions of Molly e.g., the probability that Molly will increase her speed in the next

seconds. It should also be able to provide information about the considered plans (including the chosen ones) and the risk values associated with the considered plans.

Suggestion: We suggest that when requested by authorised entities, AVs should be able to provide real-time data access (both in-vehicle or remotely) to the levels of uncertainty associated with its current and possible future digital representations of its environment and the uncertainty threshold upon which a plan or a risk mitigation action was selected.

Action This deals with the provision of information to provide insight into the execution of the AV's plans for given contexts. It is the resulting vehicle dynamics to continuous control inputs in response to circumstances and situations observed. They are measurable outputs of the perception and decision-making steps that provide valuable insights into driving behaviour and risk. As such, the continual monitoring of actions can be used for assessing the behaviour of the AV in given circumstances.

Suggestion: We suggest that AVs should be able to provide real-time data access (both in-vehicle or remotely from the vehicle) to the actions of the AV, with respect to observations and knowledge; and the resultant decisions made, when requested by authorised entities.

Process-based Explainability

Process-based explainability in an AV is concerned with the provision of information that facilitates the independent assessment of the entire operations and governance of the AV. Process-based explainability takes perception, decision, and action data, including the governance processes of the entire AV operation. This makes it possible to reconstruct an event or accident immediately after it happens, significantly reducing the time to provide recommendations for future improvements.

Process explainability can be useful for fairness, safety and performance assessment, accountability and responsibility, and impact assessment. For fairness, the explanation outlines the steps taken across the design and the implementation of the AV to ensure that its decisions are generally unbiased and fair and whether someone has been treated equitably. For example, whether adequate measures

were implemented to provide PRMs more time to cross the road; whether the pedestrian detection/classification system works with the same accuracy for people of colour; or whether predicted crash outcomes are representative of all members of the population, e.g., all body types, not just typical adult males.

For safety and performance evaluation, the explanation provides information on the steps taken across the design and implementation of the AV to maximise the accuracy, reliability, security, and robustness of its decisions and behaviours. For accountability and responsibility, the explanation provides insight into who is involved in the development and management of an AV, and who to contact for a human review of a decision. For impact assessment, the explanation provides information on the steps taken across the design and implementation of AV technologies. It considers and monitors the impacts that the use of the AV and its decisions has or may have on an individual and on broader society.

Suggestion: We suggest that real-time access to perception, decision, and action data and information about the process management (both in-vehicle or remotely from the vehicle) be made available for independent real-time processing to authorised authorities. Powers should also be granted to relevant authorities to impose sanctions in real-time based upon the failure to meet explainability requirements.

7.5 Epilogue

Vehicular means of transportation have evolved through human civilisation, from the use of chariots to highly automated vehicles. This means a reduction in physical control, a reduction in travelling time, and an increase in comfort. While an excellent opportunity, the introduction of highly automated vehicles poses concerns related to safety and societal trust due to the sophisticated nature of this new technology. Does this mean that autonomous vehicles in themselves are a threat? Does this mean that we are losing control over our automobiles? While these are not entirely true, they are questions to address if autonomous vehicles must be widely adopted in society.

First, how do consumers and regulators practically verify AVs' safety? One way to approach the assessment of perceived safety—as we have argued in this thesis—is by making the AV provide intelligible explanations with causal attributions. The frequent use of an explainer in an AV can help in-vehicle passengers assess the safety of the vehicle and also correctly calibrate their trust. For regulators, more comprehensive and detailed explanations are required. Such comprehensive explanations from driving episodes could be aligned with the corresponding scenes for which they were generated, and an after-the-fact evaluation of the vehicle's safety can be done.

An important question is whether these explanations would serve their intended purposes. While this question is quite futuristic, we investigated the effect of natural language explanations targeted at AV passengers through a physical driving simulator and a state-of-the-art virtual reality headset. It turned out that the imperfection of the AV perception system, whose output is an input to our explanation generation algorithms, led to a degraded passengers' perception of safety and an increased feeling of anxiety. Especially when these imperfections or errors were exposed by the explanations. There is a trade-off between high transparency and the perception of safety. With high transparency, some inconsequential errors in the AV might be revealed to passengers, which in turn, would increase the feeling of anxiety and even lead to a degraded perception of safety. However, we infer from the results of our studies that passengers would prefer highly transparent AVs that provide specific explanations to non-transparent AVs which provide very vague or abstract explanations.

While research has shown that explanations are helpful in autonomous driving, at least in increasing end-users' understanding of the AV's behaviour, some argue that explainability in AVs may be unnecessary when policies/regulations are made that would lead to the prevalence of AVs in society. In such situations, the pervasiveness of AVs would reduce peoples' hesitations about using AVs. It is quite unclear what the future of AVs looks like, as regulations and policies governing AVs in this early stage are evolving; some are in favour of the deployment of AVs on public roads

while others are yet to. Whatever the case may be, it is our expectation that this work acts as a catalyst to inspire future research and provides a foundation upon which to build effective, transparent and explainable autonomous vehicles.

Appendices

A

Supplementary Materials

Contents

A.0.1 Tree SHAP and Local Increments Comparison Result

Table A.1: Comparison of Tree SHAP and Local Increments contextual importance estimation methods using BLEU-4 metric. RLC: Right lane change; LLC: Left lane change.

	Tree SHAP	**Local Increment**
Stop	0.605	0.587
Move	0.615	0.236
RLC	0.471	0.228
LLC	0.533	0.407

Table A.1 shows the results of the comparative test performed to assess the performance of Tree SHAP and Local Increments contextual importance estimation methods. Each of the two contextual importance methods was used in Algorithm 1 (the second factual explanation generation algorithm) with the SAX test dataset (provided in Chapter 5). Tree SHAP method outperformed the Local Increments

method (Table A.1). The BLEU-4 metric was used as a performance measure; higher scores are better. From Table A.1, it is clear that using Tree SHAP in estimating contextual importance results in more intelligible explanations compared to the Local Increments method.

A.0.2 Snapshots from Dataset and Experiment Scenes

Figure A.1 shows snapshots of different scene varieties from the SAX dataset. Scenes include stops actions, move actions, left and right lane change actions of the ego vehicle at different road structures e.g., intersection, dual lane and single lanes. Dataset is also diverse in terms of the time of the day. Some were collected during the day while some were collected in the evening with limited light.

Figure A.1: SAX dataset is made up of diverse driving scenes at different times of the day (day and evening.) ped. - pedestrian, veh. - vehicle, em. - emergency, Mov. - moving, CLL - changes lane to the left, CLR - changes lane to the right

Figure A.2: SAX data collection field trial in London, April 2021

A. Supplementary Materials

A screenshot of the demo showcased at the 2022 Goodwood Festival of Speed is shown in Figure A.3. The explanation generation algorithms were deployed in Carla simulator (version 0.9.13). This algorithm generates explanations when the ego vehicle in Carla is driven either in autopilot or manual mode.

Figure A.3: Automated commentary driving at Goodwood festival of speed, June 2022.

A screenshot of a subset of the participants in the lab experiment on the assessment of the effects of explanations granularity and AV perception system errors is shown in Figure A.4. Participants completed a set of questionnaires and participated in manual and autonomous driving exercises in highly immersive mode using a state-of-the-art virtual reality headset and physical driving simulator.

Figure A.4: Explanation granularity and AV perception system errors experiment with virtual reality and physical driving simulator.

References

A right to explanation [Accessed: Jul. 24, 2020]. (n.d.). *The Alan Turing Institute*. https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation

Abby, N. (2020). Men or Women – Who's the Safest Drivers? [Accessed: Sept. 1, 2020]. https://carleasespecialoffers.co.uk/blog/men-or-women-whos-the-safest-drivers/

Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., & Coughlin, J. F. (2016). Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences. *Massachusetts Inst. Technol, AgeLab, Cambridge*, *1*, 16.

Acuna, D., Philion, J., & Fidler, S. (2021). Towards Optimal Strategies for Training Self-Driving Perception Models in Simulation. *Advances in Neural Information Processing Systems*, *34*, 1686–1699.

Adadi, A., & Berrada, M. (2018a). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

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

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.

Ai, Q., & Narayanan. R, L. (2021). Model-agnostic vs. model-intrinsic interpretability for explainable product search. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 5–15.

Akula, A. R., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J. Y., & Zhu, S.-C. (2019). X-ToM: explaining with theory-of-mind for gaining justified human trust. *arXiv preprint arXiv:1909.06907*.

An overview of taxonomy, legislation, regulations, and standards for automated mobility [Accessed: Feb. 16, 2020]. (2020). *Apex.AI*. https://www.apex.ai/post/legislation-standards-taxonomy-overview

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Montreal, Canada*, 1078–1088.

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

Asha, A. Z., Smith, C., Oehlberg, L., Somanath, S., & Sharlin, E. (2020). Views from the Wheelchair: Understanding Interaction between Autonomous Vehicle and Pedestrians with Reduced Mobility. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–8.

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour*, *18*, 47–57.

Ben-Younes, H., Zablocki, É., Pérez, P., & Cord, M. (2020). Driving Behavior Explanation with Multi-level Fusion. *arXiv preprint arXiv:2012.04983*.

Bidot, J., Biundo, S., Heinroth, T., Minker, W., Nothdurft, F., & Schattenberg, B. (2010). Verbal Plan Explanations for Hybrid Planning. *MKWI*, 2309–2320.

Bilgic, M., & Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. promotion. *Beyond Personalization Workshop, IUI*, *5*, 153.

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*(4), 1468.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173–189.

Biundo, S., & Schattenberg, B. (2014). From abstract crisis to concrete relief—a preliminary report on combining state abstraction and htn planning. *Sixth European Conference on Planning*.

Böhm, K., Kubjatko, T., Paula, D., & Schweiger, H.-G. (2020). New developments on EDR (Event Data Recorder) for automated vehicles. *Open Engineering*, *10*(1), 140–146.

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Ackel, L. J., Muller, U., Yeres, P., & Zieba, K. (2018). Visualbackprop: Efficient visualization of CNNs for autonomous driving. *IEEE International Conference on Robotics and Automation (ICRA)*, 4701–4708.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Bonnefon, J.-F., Černy, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., Martens, M., Mladenovic, M., Palade, P., Reed, N., et al. (2020). Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility.

Borgo, R., Cashmore, M., & Magazzeni, D. (2018). Towards providing explanations for AI planner decisions. *arXiv preprint arXiv:1810.06338*.

Bose, U. (2014). The black box solution to autonomous liability. *Wash. UL Rev.*, *92*, 1325.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal

dataset for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Chakraborti, T., Kambhampati, S., Scheutz, M., & Zhang, Y. (2017). AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*.

Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D. E., & Kambhampati, S. (2019). Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. *Proceedings of the International Conference on Automated Planning and Scheduling*, *29*, 86–96.

Chakraborti, T., Sreedharan, S., Grover, S., & Kambhampati, S. (2019). Plan Explanations as Model Reconciliation–An Empirical Study. *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 258–266.

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020a). The Emerging Landscape of Explainable Automated Planning & Decision Making.

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020b). The Emerging Landscape of Explainable Automated Planning & Decision Making. *IJCAI*, 4803–4811.

Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*.

Chang, S., Harper, F. M., & Terveen, L. G. (2016). Crowd-based personalized natural language explanations for recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 175–182.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847.

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*(2), e86474.

Chen, S., Dong, J., Du, R., Li, Y., & Labi, S. (2021). Reason induced visual attention for explainable autonomous driving. *arXiv preprint arXiv:2110.07380*.

Ciatto, G., Schumacher, M. I., Omicini, A., & Calvaresi, D. (2020). Agent-based explanations in AI: Towards an abstract framework. *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 3–20.

Clamann, M., Aubert, M., & Cummings, M. L. (2017). *Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles* (tech. rep.).

Cleger, S., Fernández-Luna, J. M., & Huete, J. F. (2014). Learning from explanations in recommender systems. *Information Sciences*, *287*, 90–108.

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. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 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. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Colley, M., Krauss, S., Lanzer, M., & Rukzio, E. (2021). How should Automated Vehicles Communicate Critical Situations? A Comparative Analysis of Visualization Concepts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(3), 1–23.

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*(2), 1–21.

Collingwood, L. (2017). Privacy implications and liability issues of autonomous vehicles. *Information & Communications Technology Law*, *26*(1), 32–45.

Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition*, *10*(5), 496–502.

Cultrera, L., Seidenari, L., Becattini, F., Pala, P., & Del Bimbo, A. (2020). Explaining autonomous driving by learning end-to-end visual attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 340–341.

Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv preprint arXiv:2006.11371*.

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*(4), 553.

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series*.

Demmel, S., Gruyer, D., Burkhardt, J.-M., Glaser, S., Larue, G., Orfila, O., & Rakotonirainy, A. (2019). Global risk assessment in an autonomous driving context: Impact on both the car and the driver [2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018]. *IFAC-PapersOnLine*, *51*(34), 390–395. https://doi.org/https://doi.org/10.1016/j.ifacol.2019.01.009

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. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Dodwell, P. (1960). Causes of behaviour and explanation in psychology. *Mind*, 1–13.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability

of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

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

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697–718.

Ebi, K. L. (2009). Catalogue of Risks: Natural, Technical, Social and Health Risks.

Eby, D. W., Molnar, L. J., Zhang, L., Louis, R. M. S., Zanier, N., Kostyniuk, L. P., & Stanciu, S. (2016). Use, perceptions, and benefits of automotive technologies among aging drivers. *Injury epidemiology*, *3*(1), 28.

Edwards, C. (2014). Car safety with a digital dashboard. *Engineering & Technology*, *10*(9), 60–64.

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. *Extended abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. (2021). Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9710–9719.

Faas, S. M., Mattes, S., Kao, A. C., & Baumann, M. (2020). Efficient paradigm to measure street-crossing onset time of pedestrians in video-based interactions with vehicles. *Information*, *11*(7), 360.

Frith, C., & Frith, U. (2005). Theory of mind. *Current biology*, *15*(17), R644–R645.

Fu, E., Johns, M., Hyde, D. A., Sibi, S., Fischer, M., & Sirkin, D. (2020). Is Too Much System Caution Counterproductive? Effects of Varying Sensitivity and Automation Levels in Vehicle Collision Avoidance Systems. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, *72*(4), 367–382.

Gibaldi, A., & Sabatini, S. P. (2021). The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis. *Behavior Research Methods*, *53*(1), 167–187.

Gillmore, S. C., & Tenhundfeld, N. L. (2020). The good, the bad, and the ugly: Evaluating tesla's human factors in the wild west of self-driving cars. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 67–71.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*(1), 39–60.

GOV.UK. (n.d.). Ethics, Transparency and Accountability Framework for Automated Decision-Making [Accessed: Jul, 2021]. %7Bhttps://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making%20%7D

Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. *R J.*, *9*(1), 421.

Guesmi, M., Chatti, M. A., Vorgerd, L., Joarder, S., Zumor, S., Sun, Y., Ji, F., & Muslim, A. (2021). On-demand personalized explanation for transparent recommendation. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 246–252.

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*(5), 1–42.

Guo, H., Meamari, E., & Shen, C.-C. (2018). Blockchain-inspired event recording system for autonomous vehicles. *1st IEEE International Conference on Hot Information-Centric Networking (HotICN)*, 218–222.

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.

Hada, D. V., & Shevade, S. K. (2021). ReXPlug: Explainable Recommendation using Plug-and-Play Language Model. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 81–91.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer*, *51*(9), 28–36.

Harris, D. M., & Harris, S. L. (2013). 3 - sequential logic design. In D. M. Harris & S. L. Harris (Eds.), *Digital design and computer architecture (second edition)* (Second Edition, pp. 108–171). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-394424-5.00003-3

Hart, W., Albarracın, 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*(4), 555.

Hayes, B., & Shah, J. A. (2017). Improving robot controller transparency through autonomous policy explanation. *12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 303–312.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. *European Conference on Computer Vision*, 3–19.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*.

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.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, *58*(3), 509–519.

Herlocker, J. L. (1999). Position statement-explanations in recommender systems. *Proceedings of the CHI*, *99*.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241–250.

Hewitt, C., Politis, I., Amanatidis, T., & Sarkar, A. (2019). Assessing public perception of self-driving cars: The autonomous vehicle acceptance model. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 518–527.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68–73.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hoffmann, J., & Magazzeni, D. (2019). Explainable AI Planning (XAIP): Overview and the Case of Contrastive Explanation. *Reasoning Web. Explainable Artificial Intelligence*, 277–282.

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The system causability scale (SCS). *KI-Künstliche Intelligenz*, 1–6.

Honda sustainability report (Tech. Rep.) [Accessed: Jun. 25, 2021]. (n.d.). https://global.honda/about/sustainability/report/pdf-download/2015.html

Hou, M., Mahadevan, K., Somanath, S., Sharlin, E., & Oehlberg, L. (2020). Autonomous Vehicle-Cyclist Interaction: Peril and Promise. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12.

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., & Ondruska, P. (2020). One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*.

Hussain, R., & Zeadally, S. (2018). Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys & Tutorials*, *21*(2), 1275–1313.

IEEE Standard for Transparency of Autonomous Systems. (2022). *IEEE Std 7001-2021*, 1–54. https://doi.org/10.1109/IEEESTD.2022.9726144

Information Commissioner's Office. (n.d.). What goes into an explanation? [Accessed: Jul, 2021]. 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/

Ingrand, F., & Ghallab, M. (2017). Deliberation for autonomous robots: A survey. *Artificial Intelligence*, *247*, 10–44.

Intelligent transport systems [Accessed: Jul. 24, 2020]. (2018). *ETSI*. https://www.etsi.org/images/files/ETSITechnologyLeaflets/IntelligentTransportSystems.pdf

ITU. (2020). "The Molly Problem" [Accessed July 2, 2021]. https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Janai, J., Güney, F., Behl, A., Geiger, A., et al. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision*, *12*(1–3), 1–308.

Jardim, A. S., Quartulli, A. M., & Casley, S. V. (2013). A study of public acceptance of autonomous cars. *Worcester Polytechnic Institute: Worceste r, MA, USA*.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, *4*(1), 53–71.

Jo, K., Kim, J., Kim, D., Jang, C., & Sunwoo, M. (2014). Development of autonomous car—Part I: Distributed system architecture and development process. *IEEE Transactions on Industrial Electronics*, *61*(12), 7131–7140.

Jones, G. W. (1992). The search for local accountability. *Strengthening local government in the 1990s*, 49–78.

Jung, M. F., Sirkin, D., Gür, T. M., & Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2201–2210.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

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.

Khodayari, A., Ghaffari, A., Ameli, S., & Flahatgar, J. (2010). A historical review on lateral and longitudinal control of autonomous vehicle motions. *International Conference on Mechanical and Electrical Technology*, 421–429.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *International Conference on Machine Learning*, 2668–2677.

Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. *Proceedings of the IEEE International Conference on Computer Vision*, 2942–2950.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*, 563–578.

Kohler, W. J., & Colbert-Taylor, A. (2014). Current law and potential legal issues pertaining to automated, autonomous and connected vehicles. *Santa Clara Computer & High Tech. LJ*, *31*, 99.

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*(4), 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*(4), 377–392.

Korpan, R., & Epstein, S. L. (2018). Toward Natural Explanations for a Robot's Navigation Plans. *Notes from the Explainable Robotic Systems Workshop, Human-Robot Interaction*.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized explanations for hybrid recommender systems. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 379–390.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, *62*(5), 718–736.

Kraus, J. M. (2020). *Psychological processes in the formation and calibration of trust in automation* (Doctoral dissertation). Universität Ulm.

Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, *11*(3), 1–17.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2018). Augmented Reality Displays for Communicating Uncertainty Information in Automated Driving. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 164–175.

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*(3), 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. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 329–341.

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

Lagstrom, T., & Lundgren, V. M. (2015). AVIP-Autonomous vehicles interaction with pedestrians. *Master of Science Thesis, Chalmers University of Technology*.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684.

Langley, P. (2019). Varieties of Explainable Agency. *ICAPS Workshop on Explainable AI Planning (XAIP)*.

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. *AAAI*, *17*, 4762–4763.

Langlois, S. (2013). ADAS HMI using peripheral vision. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 74–81.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1–8.

Lavrinc, D. (2016). This is how bad self-driving cars suck in rain [Accessed: Jul. 1, 2020]. https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/

Lavrinc, D. (2018). Cars with a digital dashboard [Accessed: Jul. 24, 2020]. https://www.buyacar.co.uk/cars/902/cars-with-a-digital-dashboard

Law Commission. (n.d.). Automated Vehicles: Analysis of Responses to the Preliminary Consultation Paper [Accessed: Jul, 2021]. https://www.scotlawcom.gov.uk/files/2815/6093/3787/Automated_Vehicles_-_Analysis_of_Responses_to_the_Preliminary_Consultation_Paper.pdf

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., Camara, F., Rothmueller, M., Vendelbo-Larsen, S. A., Rasmussen, P. H., et al. (2021). Road users rarely use explicit communication when interacting in today's traffic: implications for automated vehicles. *Cognition, Technology & Work*, *23*(2), 367–380.

Li, C., Chan, S. H., & Chen, Y.-T. (2020). Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference. *arXiv preprint arXiv:2003.02425*.

Li, J., Currano, R., Sirkin, D., Goedicke, D., Tennent, H., Levine, A., Evers, V., & Ju, W. (2020). On-Road and Online Studies to Investigate Beliefs and Behaviors of Netherlands, US and Mexico Pedestrians Encountering Hidden-Driver Vehicles. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 141–149.

Li, Y., & Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, *37*(4), 50–61.

Li, Z., Zhou, X., Wang, X., & Guo, Z. (2013). Study on subjective and objective safety and application of expressway. *Procedia-Social and Behavioral Sciences*, *96*, 1622–1630.

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. *Proceedings of the 11th International Conference on Ubiquitous Computing*, 195–204.

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

Liu, T., Zhou, H., Itoh, M., & Kitazaki, S. (2018). The impact of Explanation on Possibility of Hazard Detection Failure on Driver Intervention under Partial Driving Automation. *IEEE Intelligent Vehicles Symposium (IV)*, 150–155.

Liu, Y.-C., Hsieh, Y.-A., Chen, M.-H., Yang, C.-H. H., Tegner, J., & Tsai, Y.-C. J. (2020). Interpretable self-attention temporal reasoning for driving behavior understanding. *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2338–2342.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

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? *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, *49*(5), 773–785.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems*, *53*, 6–8.

Mahadevan, K., Somanath, S., & Sharlin, E. (2018). Communicating awareness and intent in autonomous vehicle-pedestrian interaction. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12.

Mahmud, S. S., Ferreira, L., Hoque, M. S., & Tavassoli, A. (2017). Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. *IATSS Research*, *41*(4), 153–163. https://doi.org/10.1016/j.iatssr.2017.02.001

Managing a Slow Reaction Time While Driving. (2019). https://drivesafety.com/managing-a-slow-reaction-time/#:~:text=Reaction%5C%20times%5C%20vary%5C%20from%5C%20one,seconds%5C%20or%5C%20even%5C%201%5C%20second

Markkula, G., Benderius, O., Wolff, K., & Wahde, M. (2012). A review of near-collision driver behavior models. *Human factors*, *54*(6), 1117–1143.

Markkula, G., Madigan, R., Nathanael, D., Portouli, E., Lee, Y. M., Dietrich, A., Billington, J., Schieben, A., & Merat, N. (2020). Defining interactions: A conceptual framework for understanding interactive behaviour in human and automated road traffic. *Theoretical Issues in Ergonomics Science*, *21*(6), 728–752.

Marques, L., Vasconcelos, V., Pedreiras, P., & Almeida, L. (2011). A flexible dashboard panel for a small electric vehicle. *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, 1–4.

Marsland, T. A. (1986). A review of game-tree pruning. *ICGA journal*, *9*(1), 3–19.

Martinesco, A., Netto, M., Neto, A. M., & Etgens, V. H. (2019). A Note on Accidents Involving Autonomous Vehicles: Interdependence of Event Data Recorder, Human-Vehicle Cooperation and Legal Aspects. *IFAC-PapersOnLine*, *51*(34), 407–410.

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews*, 1–22.

Mashko, A., Bouchner, P., Rozhdestvenskiy, D., & Novotnỳ, S. (2016). Virtual traffic signs-assessment of an alternative ADAS user interface with use of driving simulator. *Advances in Transportation Studies*, (1).

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

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, 1–11.

Michael, M., & Schlipsing, M. (2015). Extending traffic light recognition: Efficient classification of phase and pictogram. *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. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288.

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. *IEEE 18th International Conference on Intelligent Transportation Systems*, 2458–2464.

Mok, B., Sirkin, D., Sibi, S., Miller, D. B., & Ju, W. (2015). Understanding Driver-Automated Vehicle Interactions Through Wizard of Oz Design Improvisation. *Driving Assessment Conference*.

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 295–307.

Moore, D., Strack, G. E., Currano, R., & Sirkin, D. (2019). Visualizing implicit eHMI for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 475–477.

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

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, *27*(5-6), 527–539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Mulgan, R. (2000). 'Accountability': An ever-expanding concept? *Public administration*, *78*(3), 555–573.

Nahata, R., Omeiza, D., Howard, R., & Kunze, L. (2021). Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety. *24th International Conference on Intelligent Transportation Systems (ITSC)*.

Najm, W. G., Smith, J. D., Yanagisawa, M., et al. (2007). *Pre-crash scenario typology for crash avoidance research* (tech. rep.). United States. National Highway Traffic Safety Administration.

Neerincx, M. A., van der Waa, J., Kaptein, F., & van Diggelen, J. (2018). Using perceptual and cognitive explanations for enhanced human-agent team performance. *International Conference on Engineering Psychology and Cognitive Ergonomics*, 204–214.

NHTSA. (n.d.). Event Data Recorders [Accessed July 2, 2021]. https://www.nhtsa.gov/fmvss/event-data-recorders-edrs

NTBS. (2018). Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View, California [Accessed: Oct. 30, 2020]. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf

Oliveira, L., Proctor, K., Burns, C. G., & Birrell, S. (2019). Driving style: how should an automated vehicle behave? *Information*, *10*(6), 219.

Omeiza, D., Kollnig, K., Webb, H., Jirotka, M., & Kunze, L. (2021). Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving. *IEEE International Conference on Advanced Robotics and its Social Impacts*.

Omeiza, D., Speakman, S., Cintas, C., & Weldermariam, K. (2019). Smooth Grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*.

Omeiza, D., Web, H., Jirotka, M., & Kunze, L. (2021). Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. *2021 IEEE Intelligent Vehicles Symposium (IV)*.

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2022). Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865

Omicini, A. (2020). Not Just for Humans: Explanation for Agent-to-Agent Communication. *DP@ AI* IA*, 1–11.

Ozbay, K., Yang, H., Bartin, B., & Mudigonda, S. (2008). Derivation and validation of new simulation-based surrogate safety measure. *Transportation research record*, *2083*(1), 105–113.

Palazzi, A., Abati, D., Calderara, S., Solera, F., & Cucchiara, R. (2018). Predicting the Driver's Focus of Attention: the DR(eye)VE Project.

Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013). Interpreting random forest models using a feature contribution method. *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, 112–119.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, S. Y., Moore, D. J., & Sirkin, D. (2020). What a Driver Wants: User Preferences in Semi-Autonomous Vehicle Decision-Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13.

Parr, T., & Grover, P. (2020). How to visualize decision trees [Accessed: Mar. 9, 2023]. https://explained.ai/decision-tree-viz/

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human factors*, *58*(2), 229–241.

Pieters, W. (2011). Explanation and trust: what to tell the user in security and AI? *Ethics and information technology*, *13*(1), 53–64.

Pinter, K., Szalay, Z., & Vida, G. (2020). Road accident reconstruction using on-board data, especially focusing on the applicability in case of autonomous vehicles. *Periodica Polytechnica Transportation Engineering*.

Piramuthu, O. B., & Caesar, M. (2021). How Effective are Identification Technologies in Autonomous Driving Vehicles? *International Conference on Advanced Communication Technologies and Networking (CommNet)*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–52.

Prolific. (n.d.). https://www.prolific.co/

Pu, P., & Chen, L. (2006). Trust building with explanation interfaces. *Proceedings of the 11th International Conference on Intelligent user Interfaces*, 93–100.

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*(1), 28–41.

Raab, E. L. (1985). Normal saccadic velocities. *Journal of Pediatric Ophthalmology & Strabismus*, *22*(1), 20–22.

Rahimpour, A., Martin, S., Tawari, A., & Qi, H. (2019). Context Aware Road-user Importance Estimation (iCARE). *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2337–2343.

Rajaonah, B., Anceaux, F., & Vienne, F. (2006). Trust and the use of adaptive cruise control: a study of a cut-in situation. *Cognition, Technology & Work*, *8*(2), 146–155.

Raman, V., & Kress-Gazit, H. (2012). Explaining impossible high-level robot behaviors. *IEEE Transactions on Robotics*, *29*(1), 94–104.

Raman, V., Lignos, C., Finucane, C., Lee, K. C., Marcus, M. P., & Kress-Gazit, H. (2013). Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. *Robotics: Science and Systems*, *2*(1), 2–1.

Ramanishka, V., Chen, Y.-T., Misu, T., & Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7699–7707.

Rasouli, A., & Tsotsos, J. K. (2019). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems*, *21*(3), 900–918.

Reid, T. G., Houts, S. E., Cammarata, R., Mills, G., Agarwal, S., Vora, A., & Pandey, G. (2019). Localization requirements for autonomous vehicles. *arXiv preprint arXiv:1906.01061*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rizzo, S. G., Vantini, G., & Chawla, S. (2019). Reinforcement learning with explainability for traffic signal control. *IEEE Intelligent Transportation Systems Conference (ITSC)*, 3567–3572.

Roth-Berghofer, T. R. (2004). Explanations and case-based reasoning: Foundational issues. *Proceedings of the European Conference on Case-Based Reasoning*, 389–403.

Saabas, A. (2014). Interpreting random forests. http://blog.datadive.net/interpreting-random-forests/

Sado, F., Loo, C. K., Kerzel, M., & Wermter, S. (2020). Explainable Goal-Driven Agents and Robots–A Comprehensive Review and New Framework. *arXiv preprint arXiv:2004.09705*.

Samek, W., Wiegand, T., & Müller, K.-R. (2017a). EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS.

Samek, W., Wiegand, T., & Müller, K.-R. (2017b). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sasai, S., Kitahara, I., Kameda, Y., Ohta, Y., Kanbara, M., Morales, Y., Ukita, N., Hagita, N., Ikeda, T., & Shinozawa, K. (2015). MR visualization of wheel

trajectories of driving vehicle by seeing-through dashboard. *IEEE International Symposium on Mixed and Augmented Reality Workshops*, 40–46.

Schmidt, A., Dey, A. K., Kun, A. L., & Spiessl, W. (2010). Automotive user interfaces: human computer interaction in the car. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3177–3180).

Schneider, T., Hois, J., Rosenstein, A., Ghellal, S., Theofanou-Fülbier, D., & Gerlicher, A. R. (2021). ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Scottish Law Commission. (n.d.). Automated Vehicles: Consultation Paper 3 - A regulatory framework for automated vehicles [Accessed: Jun. 11, 2021]. https://www.scotlawcom.gov.uk/files/3916/0854/6794/AV-CP3-18-12-20.pdf

Selkowitz, A. R., Larios, C. A., Lakhmani, S. G., & Chen, J. Y. (2017). Displaying information to support transparency for autonomous platforms. In *Advances in Human Factors in Robots and Unmanned Systems* (pp. 161–173). Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shalev-Shwartz, S., & Ben-David, S. (2014). Decision trees. understanding machine learning.

Shashua, A., & Shalev-Shwartz, S. (2017). A plan to develop safe autonomous vehicles. And prove it. *Intel Newsroom*, 8.

Shen, Y., Jiang, S., Chen, Y., Yang, E., Jin, X., Fan, Y., & Campbell, K. D. (2020). To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles. *arXiv preprint arXiv:2006.11684*.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Silveira, M. S., de Souza, C. S., & Barbosa, S. D. (2001). Semiotic engineering contributions for designing online help systems. *Proceedings of the 19th Annual International Conference on Computer Documentation*, 31–38.

Silvera, G., Biswas, A., & Admoni, H. (2022). DReyeVR: Democratizing Virtual Reality Driving Simulation for Behavioural & Interaction Research. *ACM/IEEE Human Robot Interaction Conference*, 639–643.

Singh, G., Akrigg, S., Di Maio, M., Fontana, V., Alitappeh, R. J., Saha, S., Jeddisaravi, K., Yousefi, F., Culley, J., Nicholson, T., et al. (2021). Road: The road event awareness dataset for autonomous driving. *arXiv preprint arXiv:2102.11585*.

Sippy, J., Bansal, G., & Weld, D. S. (2020). Data staining: A method for comparing faithfulness of explainers. *Proc. of ICML Workshop on Human Interpretability in Machine Learning (WHI)*.

Sirkin, D., Martelaro, N., Johns, M., & Ju, W. (2017). Toward measurement of situation awareness in autonomous vehicles. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 405–415.

Smith, B. W., & Svensson, J. (2015). Automated and autonomous driving: regulation under uncertainty.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Sreedharan, S., Chakraborti, T., & Kambhampati, S. (2017). Balancing explicability and explanation in human-aware planning. *2017 AAAI Fall Symposium*, 61–68.

Sreedharan, S., Srivastava, S., Smith, D., & Kambhampati, S. (2019). Why Can't You Do That HAL? Explaining Unsolvability of Planning Tasks. *International Joint Conference on Artificial Intelligence*.

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.

Stepin, I., Catala, A., Pereira-Fariña, M., & Alonso, J. M. (2021). Factual and Counterfactual Explanation of Fuzzy Information Granules. *Interpretable Artificial Intelligence: A Perspective of Granular Computing*, *937*, 153.

Šucha, M. (2014). Road users' strategies and communication: driver-pedestrian interaction. *Transport Research Arena (TRA)*.

Sun, D., Ukkusuri, S., Benekohal, R. F., & Waller, S. T. (2003). Modeling of motorist-pedestrian interaction at uncontrolled mid-block crosswalks. *Transportation Research Record, TRB Annual Meeting CD-ROM, Washington, DC*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.

Sustainable value report (Tech. Rep.) [Accessed: Jun. 25, 2021]. (n.d.). https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/ir/downloads/en/2016/2016-BMW-Group-Sustainable-Value-Report.pdf

Suthaharan, S. (2016). Decision Tree Learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, 237–269.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., & Muhammad, N. (2020). A survey of end-to-end driving: Architectures and training methods. *IEEE Transactions on Neural Networks and Learning Systems*.

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning*, 10096–10106.

Tang, Z., Chuang, K. V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M. J., & Dugger, B. N. (2019). Interpretable classification of Alzheimer's disease

pathologies with a convolutional neural network pipeline. *Nature communications*, *10*(1), 1–14.

Terken, J., & Pfleging, B. (2020). Toward shared control between automated vehicles and users. *Automotive Innovation*, *3*(1), 53–61.

Tesla deaths [Accessed: Jul. 24, 2021]. (n.d.). *TeslaDeaths.com*. https://www.tesladeaths.com/

TEuropean Commission - Press release: Road safety: Commission welcomes agreement on new EU rules to help save lives [Accessed: Feb. 8, 2021]. (2019). https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1793

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

Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI): towards medical XAI. *arXiv preprint arXiv:1907.07374*.

TomTom launches map-based ADAS software platform Virtual Horizon [Accessed: Aug. 10, 2021]. (n.d.). *TomTom*. https://safecarnews.com/toyota-launches-map-based-adas-software-platform-virtual-horizon/

UNECE. (2019). Working Party on Automated/Autonomous and Connected Vehicles (GRVA): EDR/DSSAD 1st session. EDR-DSSAD-01-06 Overview of EDR [Accessed: Feb. 8, 2021]. https://wiki.unece.org/pages/viewpage.action?pageId=87621710

van der Linden, I., Haned, H., & Kanoulas, E. (2019). Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039*.

Voigt, P., & Von 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. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–15.

Wang, J., Zhang, L., Huang, Y., & Zhao, J. (2020). Safety of autonomous vehicles. *Journal of Advanced Transportation*, *2020*.

Wang, L., Zhang, Y., & Wang, J. (2017). Map-based localization method for autonomous vehicles using 3D-LIDAR. *IFAC-PapersOnLine*, *50*(1), 276–281.

Ward, J. R., Agamennoni, G., Worrall, S., Bender, A., & Nebot, E. (2015). Extending time to collision for probabilistic reasoning in general traffic scenarios. *Transportation Research Part C: Emerging Technologies*, *51*, 66–82. https://doi.org/https://doi.org/10.1016/j.trc.2014.11.002

Wiegand, G., Eiband, M., Haubelt, M., & Hussmann, H. (2020). "I'd like an Explanation for That!" Exploring Reactions to Unexpected Autonomous Driving. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–11.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wilde, G. S. (1980). Immediate and delayed social interaction in road user behaviour. *Applied Psychology*, *29*(4), 439–460.

Wilson, S. J., Glue, P., Ball, D., & Nutt, D. J. (1993). Saccadic eye movement parameters in normal subjects. *Electroencephalography and Clinical Neurophysiology*, *86*(1), 69–74.

Wintersberger, P., Nicklas, H., Martlbauer, T., Hammer, S., & Riener, A. (2020). Explainable automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 252–261.

World Report for Intelligent Transport Systems (ITS) Standards - A Joint APEC-International Organization for Standardization (ISO) Study of Progress to Develop and Deploy ITS Standards (ISO TR 28682) [Accessed: July 24, 2020]. (2017). *Asian-Pacific Economic Cooperation*. https://apec.org/Publications

Wu, B.-F., Chen, Y.-H., & Yeh, C.-H. (2013). Driving behaviour-based event data recorder. *IET Intelligent Transport Systems*, *8*(4), 361–367.

Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., & Vasconcelos, N. (2020). Explainable Object-induced Action Decision for Autonomous Vehicles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, J., & Coughlin, J. F. (2014). In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers. *International Journal of Automotive Technology*, *15*(2), 333–340.

Yao, Y., & Atkins, E. (2020). The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Transactions on Intelligent Transportation Systems*.

Yao, Y., Wang, X., Xu, M., Pu, Z., Atkins, E., & Crandall, D. (2020). When, where, and what? A new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044*.

You, T., & Han, B. (2020). Traffic Accident Benchmark for Causality Recognition. *European Conference on Computer Vision*, 540–556.

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020a). A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, *8*, 58443–58469.

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020b). A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, *8*, 58443–58469.

Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2021). Explainability of vision-based autonomous driving systems: Review and challenges. *arXiv preprint arXiv:2101.05307*.

Zalta, E. N., Nodelman, U., Allen, C., & Perry, J. (1995). Stanford encyclopedia of philosophy.

Zang, S., Ding, M., Smith, D., Tyler, P., Rakotoarivelo, T., & Kaafar, M. A. (2019). The impact of adverse weather conditions on autonomous vehicles: How

rain, snow, fog, and hail affect the performance of a self-driving car. *IEEE Vehicular Technology Magazine*, *14*(2), 103–111.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833.

Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., & Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. *IEEE International Conference on Robotics and Automation (ICRA)*, 1313–1320.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Zhou, Y., & Danks, D. (2020). Different "Intelligibility" for Different Folks. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 194–199.

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. *IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8.