Development of Explainable Artificial Intelligence Approaches for Autonomous Vehicles

by

Shahin Atakishiyev

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

Autonomous driving, as a rapidly growing field, has received increasing attention from the general society and the automotive industry over the last two decades. However, road accidents involving autonomous vehicles have hindered societal acceptance and deployment of this technology on roads. As self-driving decisions are powered by artificial intelligence approaches, intelligent driving systems must justify their actions, particularly in critical traffic scenarios. Consequently, explainability of autonomous driving has emerged as a vital research direction in the field.

This dissertation aims to develop explainable artificial intelligence techniques for autonomous vehicles by approaching the existing issues from three essential aspects: interactivity, robustness analysis, and time granularity of explanations. In this sense, I first present a comprehensive overview of explainable artificial intelligence approaches for autonomous vehicles and describe the research gaps in this direction. Second, I introduce a visual question answering approach to explain autonomous driving actions in an interactive manner. Third, I propose a situation awareness framework for autonomous vehicles backed by explanations and human-machine interfaces. Finally, I thoroughly investigate safety implications of explainable artificial intelligence in end-to-end autonomous driving via critical case studies and an empirical analysis. Overall, in pursuit of developing explainable artificial intelligence approaches for autonomous vehicles, this dissertation highlights (1) how to build intelligible and interactive explanations, (2) critical challenges in building trustworthy interactive explanations, and (3) how to leverage explanations in enhancing self-driving safety.

Preface

This dissertation is the outcome of my doctoral research. It features content based on the following articles published or submitted for publication:

1. Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, Randy Goebel. Towards Safe, Explainable, and Regulated Autonomous Driving. In *Explainable AI for Intelligent Transportation Systems, pp. 32-52, 2023.* [DOI]

2. Shahin Atakishiyev, Mohammad Salameh, Housam Babiker, Randy Goebel. Explaining Autonomous Driving Actions with Visual Question Answering. In *Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems, Bilbao, Spain, pp. 1207-1214, 2023.* [DOI]

3. Shahin Atakishiyev, Mohammad Salameh, Randy Goebel. Incorporating Explanations into Human-Machine Interfaces for Trust and Situation Awareness in Autonomous Vehicles. *In Proceedings of the 2024 IEEE 35th Intelligent Vehicles Symposium, Jeju, South Korea, pp. 2948-2955, 2024.* [DOI]

4. Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, Randy Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *IEEE Access, Vol. 12, pp. 101603-101625, 2024* [DOI]

5. Shahin Atakishiyev, Mohammad Salameh, Randy Goebel. Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving. *Under review in IEEE Transactions on Intelligent Transportation Systems* [PDF]

"A scholar's primary aim is to make a contribution to mankind, not to win the Nobel Prize. The pleasure of such a contribution is beyond a measure..."

- Prof. Aziz Sancar,
Nobel Laureate in Chemistry in 2015.

Acknowledgments

There are many people to whom I feel indebted for their support by any means during my five-year doctoral study. First and foremost, I am fortunate to work under the supervision of Prof. Randy Goebel. Randy is a serious critic and careful scientist, yet incredibly supportive of his students. I have learned many things from him regarding research and all aspects of life. Completing my doctoral study would have been impossible without his patience, support, and supervision.

My special thanks go to Dr. Mohammad Salameh for co-supervising me over the last three years of my study. Mohammad has always given constructive feedback on my research, and is a great mentor and friend.

I am grateful to Dr. Hengshuai Yao for co-supervising me for a year when I started working on this research project. When I was a novice in reinforcement learning, autonomous driving, and explainable artificial intelligence, Hengshuai helped me understand the fundamentals of the mentioned topics together with Randy. The ideas emerging from our regular meetings and discussions shaped my research direction on this topic to a great extent. Thanks for your collaboration, Hengshuai.

I would also like to extend my gratitude to my Examining Committee, Dr. Matthew Guzdial, Dr. Ehsan Hashemi, Dr. Zhi-Jun (Tony) Qiu, Dr. Lei Ma, and Dr. Luis Miguel Bergasa, for their time, insightful discussions, and constructive feedback on my dissertation.

My biggest appreciation goes to my spouse and our lovely two-year-old son, Aykhan, who brought joy to our lives. I also apologize to them for sometimes not spending enough time with them amid meeting deadlines, working at the weekends, and overnights. Their incredible support has helped me become more motivated and a family person amid many difficulties. They have made this journey more meaningful.

I am profoundly grateful to my family in my home country, Azerbaijan, for their endless support and love. My mother, father, brother, sister, and relatives, along with their families, have always been my pillars of strength during difficult times. Particularly, my mother and father have made significant sacrifices for my education since childhood, and I can never forget their invaluable support.

I acknowledge the support of my secondary school teachers, neighbors, and everyone in my beloved village, Mirzabayli, in Gabala, Azerbaijan. Every year, when I visit back home, childhood memories come back, and I am fond of remembering cherished memories I had in my childhood with them in my lovely village.

Throughout my graduate studies, I have met many wonderful people at the University of Alberta, in Edmonton, and in my home country, Azerbaijan. I have been amazed by how correct the phrase "Assume you can learn at least one thing from every individual you meet, and you will." is. Indeed, each person is valuable with their opinions, and I have learned many things from my friends and colleagues.

I would also like to extend my gratitude to the Azerbaijani Community in Edmonton for our cultural events over the past years. Thank you for your friendship, support, and the time spent together.

I am also grateful to the Ministry of Science and Education of the Republic of Azerbaijan for their generous support for my study.

My doctoral study has been funded by the Alberta Machine Intelligence Institute (Amii), the Computing Science Department of the University of Alberta, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Shahin Atakishiyev
Edmonton, Canada
September 2024

Table of Contents

List of Tables

List of Figures

Abbreviations

ADAS Advanced Driving Assistance System.

AI Artificial Intelligence.

ALVINN Autonomous Land Vehicle In a Neural Network.

ASIL Automotive Software Integrity Level.

AV Autonomous Vehicle.

AVAM Autonomous Vehicle Acceptance Model.

BC Behavior Cloning.

BDD Berkeley DeepDrive.

BDD-OIA Berkeley DeepDrive Object-Induced Action.

BDD-A Berkeley DeepDrive Attention.

BDD-X Berkeley DeepDrive-X.

BERT Bidirectional Encoder Representations from Transformers.

CAM Class Activation Map.

CLIP Contrastive Language–Image Pre-training.

CNN Convolutional Neural Network.

DDPG Deep Deterministic Policy Gradient.

DVE Deep Visual Explanation.

ECU Electronic Control Unit.

EEC End-to-End Control.

EU European Union.

FCM Functional Causal Model.

FoV Field of View.

FSM Finite-State Machine.

GAN Generative Adversarial Network.

GCNN Graph Convolutional Neural Network.

GDPR General Data Protection Regulation.

GPS Global Positioning System.

GPT Generative Pretrained Transformer.

GPU Graphics Processing Unit.

GRIT Goal Recognition with Interpretable Tree.

HIIL Hierarchical Interpretable Imitation Learning.

HMI Human-Machine Interface.

HUD Head-Up Display.

IL Imitation Learning.

IMU Intertial Measurement Unit.

IOC Inverse Optimal Control.

IRL Inverse Reinforcement Learning.

ISO International Organization for Standardization.

LIDAR Light Detection And Ranging.

LLaVA Large Language and Vision Assistant.

LLM Large Language Model.

LSTM Long Short-Term Memory.

MDP Markov Decision Process.

MILE Model-Based Imitation Learning.

ML Machine Learning.

NACTO National Association of City Transportation Officials.

NHTSA National Highway Traffic Safety Administration.

NLP Natural Language Processing.

OCTET Object-aware Counterfactual Explanations.

POMDP Partially Observable Markov Decision Process.

QA Question-Answer.

RADAR Radio Detection And Ranging.

ReLU Rectified Linear Unit.

RL Reinforcement Learning.

RNN Recurrent Neural Networks.

SAE Society of Automotive Engineers.

SHAP Shapley Additive Explanations.

SRC Safety-Regulatory Compliance.

STEEX Steering Counterfactual Explanations with Semantics.

STL Signal Temporal Logic.

TOR Takeover Request.

TPU Tensor Processing Unit.

UX User Interface.

V2V Vehicle-to-Vehicle.

V2X Vehicle-to-Everything.

VANET Vehicular Adhoc Networks.

VGG Visual Geometry Group.

VideoQA Video Question Answering.

ViT Vision Transformer.

VLM Vision-Language Model.

VQA Visual Question Answering.

XAI Explainable Artificial Intelligence.

XRL Explainable Reinforcement Learning.

Chapter 1

Introduction

Over the last two decades, the automotive industry has seen substantial developments in the transition from manually operated vehicles to vehicles with varying levels of automation. The DARPA Grand Challenge [261] and advancements in powerful artificial intelligence (AI) approaches, particularly in deep learning and computer vision algorithms, have enabled the emergence of autonomous vehicles (AVs) and shaped a new era in intelligent transportation systems (ITS). The potential benefits of AVs are improved operational safety [182], reduced CO2 emissions [216], diminished transportation costs [1], and reduced traffic density [93]. Intel's report on the projected benefits of AVs estimates that deployment of such vehicles on roads will result in a reduction of 250 million hours of users' commuting time per year and save more than half a million lives from 2035 to 2045, just in the USA [156].

While two decades of progress and the potential impacts and benefits of AVs in everyday life are promising, there is a major societal concern about the functional safety of such vehicles. This issue, as a major drawback, originates mainly from reports of recent traffic accidents with the presence of AVs, primarily owing to their opaque decision-making [3, 87, 250, 294]. The issue is even exacerbated when the actual reason leading to such road mishaps remains unknown. Moreover, except for the safety issue, the "black-box" decision-making process also hinders resolving legal culpability and accountability issues, as a consequence [38]. In this regard, explainable AI (XAI)

has emerged as a paradigm shift in the ongoing design and development of AVs and necessitates a transparent intelligent driving system for this technology [3, 6, 16, 230]. Based on these critical issues, there are three crucial requirements for the successful adoption of modern AVs by general society: safety, explainability, and regulatory compliance [21]. The *safety* criterion refers to the acceptable and reliable performance of AVs in interaction with stationary and dynamic objects in dynamic environments and behaving appropriately in unpredictable driving conditions. The *explainability*[1] theme concentrates on providing insights into real-time action decisions of AVs (e.g., Figure 1.1). Finally, *regulatory compliance* ensures that AVs, as a holistic system, behave under traffic rules by meeting established safety standards and can justify their temporal decision-making process as required by transportation jurisdictions.

However, it is also noteworthy to underscore that providing acceptable and faithful explanations to relevant interaction partners in autonomous driving is considerably challenging. First, people have various technical knowledge and backgrounds in AV technology, and explanation format, content, and conveyance must consider this nuance. Moreover, people's various physical and cognitive abilities necessitate the design of various types of automotive human-machine interfaces (HMIs) to ensure that explanations are communicated considering individuals' needs [15, 234, 258]. Furthermore, as AV decisions are temporal, explanations must be delivered to relevant interaction partners in a timely manner. Finally, explanations must be robust, context-aware, and faithful, ensuring that they defend against potential adversarial attacks and interactions. Overall, as people with various backgrounds and abilities are targets of explanations, the construction mechanism of explanations must be human-centered.

1.1 Thesis Statement

The primary goal of this thesis is to develop XAI approaches that overcome fundamental problems of explanations for AVs — human interpretability, robustness,

[1] I will use the terms *explainability* and *interpretability* interchangeably.

Figure 1.1: A canonical example of explanation conveyance in autonomous driving: An AV presents a live natural language explanation of its real-time action to bystanders. The image has been adapted and modified from the original source: [61].

and time granularity. Throughout the following chapters, I focus on presenting principled and appropriate solutions toward this goal. In this sense, I explain the crucial aspects of each problem in this section.

1.1.1 Human Interpretability of Explanations

AV explanations can be different in their form, content, and modality depending on the recipients [20, 243]. For instance, ordinary passengers may be satisfied with a simple, linguistic explanation of AV actions. However, a system engineer and an AI scientist may need more informative and technically-rich explanations to understand the current functionalities of the car, with the motivation to appropriately "debug" the existing driving system as required. Furthermore, individuals may require different modalities of explanation communication, such as via vibrotactile, light, and sound methods [236, 237]. Consequently, depending on stakeholders' needs and preferences, AV explanations must be interpretable and intelligible for targeted interaction partners [15, 20, 200]. In this regard, human-interpretability of explanations can be analyzed within the following research questions:

What are appropriate and effective explanation delivery techniques with human factors consideration? What are the potential challenges of delivering human-interpretable explanations to the general public?

1.1.2 Robustness of Explanations

Various studies have shown that machine learning (ML) explanations are fragile to simple adversarial interactions or perturbation to model inputs [100, 280, 283]. As the primary goal of explanations is to ensure user trust and disclose the decision-making rationale of autonomous systems, the robustness of explanations is of paramount importance in achieving this goal. Particularly, when humans directly interact with such a system, the explanation model must understand conventional and adversarial interactions and thereby present context-aware and faithful explanations.

In the context of autonomous driving, the robustness of explanations has significant implications for user trust and safety. Consider an AV equipped with a conversational user interface that enables passengers on board to ask questions to its HMI about the vehicle's present action or traffic situation. Assume in an actual *right turn* scenario under a green light, a passenger asks the conversational user interface, "Why is the car turning to the *left*?" as an adversarial question. If the relevant HMI presents a response like " The car is turning to the *left* because...," such an example shows the issue with the weakness of the explanation model to detect tricky questions. On the other hand, as a desideratum, the robust model can present an answer such as "No, the car is turning to the right as the traffic light allows a right turn," explaining what the AV is doing and why it is doing so, denoting causal attribution. These questions may be asked deliberately (i.e., to stress test the explanation model) or unintentionally (i.e., people with visual impairments may have difficulty perceiving the traffic scene correctly) to test the trustworthiness of the vehicle, its action decisions, and its awareness of the operational surroundings. Consequently, automotive HMIs should not only provide conventional explanations but also defend against potential

adversarial queries to ensure that explanations are robust, reflecting the vehicle's action decision-making process. Finally, from a safety perspective, robustness against human adversarial questions can contribute to the development of safety and security measures to detect, identify, and mitigate potential adversarial attacks on AVs. Hence, the robustness perspective of explanations can be answered in the following research questions:

> *Why is there a need for robust AV explanations? What are the implications of the robustness of explanations for relevant interaction partners?*

1.1.3 Time Granularity of Explanations

Except for the contents and forms of explanations, the timing mechanism of explanations also deserves significant attention. As action decisions of AVs are time-critical, relevant explanations, depending on the target recipients, must be delivered in the allotted time interval. In general, the timing sensitivity of AV explanations can be analyzed by answering the following three research questions:

> *Should explanations be delivered before action is chosen or after action is performed? What is the appropriate lead time for a safe transition from an automatic mode to a human takeover? Should explanations be communicated to humans seamlessly or only in critical moments?*

1.2 Key Contributions

The main goal of this dissertation is to present foundations of XAI for autonomous driving in the realm of safety, transparency, and regulatory compliance backed by analytical, empirical studies, and human factors consideration. Overall, the desiderata for building such XAI approaches for AVs can be summarized as follows:
Desideratum 1: Interactive explanations that enable humans to ask questions to automotive HMI and be aware of an AV's actions and traffic situations.

Desideratum 2: Robust explanations that identify and defend against adversarial and tricky interactions.

Desideratum 3: Temporally-sensitive explanations that are delivered within the allotted time interval.

Desideratum 4: Faithful and informative explanations that can contribute to safety of autonomous driving in real time or via a retrospective analysis.

I deal with the problems from three perspectives: problem definition, data, and evaluation of the proposed solution. My contributions consider these perspectives and constitute essential factors and steps for developing explanation models for modern AV technology and for both modular and end-to-end autonomous driving. Overall, the main contributions of this dissertation can be summarized as follows:

- **Comprehensive Overview of XAI Approaches for AVs:** As explainability is becoming an integral component of autonomous driving, there is an imminent need to understand various reasons necessitating XAI-based AVs. To this end, I present a comprehensive overview and classification of XAI approaches for AVs, describing the current trends toward explainable autonomous driving systems (Chapter 3).

- **Explaining Autonomous Driving Actions with Visual Question Answering:** As a meaningful dialogue model between people on board and automotive HMI, I propose the use of a visual question answering (VQA) approach to explaining autonomous driving actions. In this sense, I train a reinforcement learning (RL) agent to control the vehicle and generate a driving video. Further, I identify scenarios within several action categories (such as go straight, turn left, turn right), convert the video to image sequences, and annotate the images with question-answer (QA) pairs and causal attributions. The goal is that once the model is trained with driving scenes and relevant QA pairs, it picks the correct answer out of a multitude of candidates to the asked question about the action taken in an unseen scene. The empirical results show that VQA is an effective approach to

enable meaningful dialogues between AV and people on board (Chapter 4).

- **Impact of Vision-Language Model-based Explanations on User Trust and Situation Awareness in AVs:** While the experiment described in Chapter 4 focuses on manually annotated data with selected action categories, real driving scenarios are usually more complex and dynamic. Hence, automotive explanation models must generalize to complicated scenarios and justify the vehicle's decisions with accurate causal reasoning. To this end, I present a general situation awareness framework backed by HMI and explanations. To validate the framework, I select critical traffic scenarios from a high-quality real-driving dataset and use a multimodal transformer for the task of VQA. Different from the methodology in Chapter 4, instead of manually annotating driving data with QA pairs, I rely on the massive internal knowledge of a pretrained multimodal transformer and test the transformer's robustness against both conventional and adversarial questions. I perform a user study in the empirical findings, show the implications of incorrect explanations in users' perceived safety and feeling of comfort with an AV and validate the findings with hypothesis testing (Chapter 6).

- **Safety Implications of Explanations:** As XAI is viewed as a promising paradigm shift to resolve trust and accountability issues, I argue that real-time and retrospective explanations can also help improve safety of autonomous driving. For this purpose, I identify critical scenarios that show the value of explanations from a safety perspective, perform an empirical study, and present a detailed analysis of the safety implications of XAI in end-to-end autonomous driving (Chapter 7).

1.3 Dissertation Outline

This dissertation is organized into nine chapters. After the introduction in Chapter 1, I present background information on AVs in terms of their design architecture and safety in Chapter 2. Chapter 3 describes a comprehensive overview of various

XAI approaches for AVs. Chapter 4 introduces a VQA framework for explaining RL-controlled autonomous driving. Chapter 5 presents a paradigm shift toward massive knowledge-based vision-language models applicable for explanation provision in AVs. Motivated by the potential limitations of the VQA framework presented in Chapter 4 and the advantage of vision-language model-based VQA described in Chapter 5, I propose a general situation awareness framework for AVs in Chapter 6 and validate it via an empirical study and human evaluation. Afterward, Chapter 7 thoroughly studies safety implications of explanations for end-to-end autonomous driving. Based on the current industrial trends in the realm of safety and explainability, Chapter 8 envisions AV2.0 and describes potential challenges and a roadmap for this goal. Finally, Chapter 9 summarizes the contributions of the dissertation and discusses potential directions for future exploration.

Chapter 2

Background

2.1 Introduction

This chapter presents background information on state-of-the-art AV technology. It revisits key aspects of autonomous driving, such as safety, fundamental issues, and regulation and standards. Afterward, the chapter describes the need for explanations for AV technology from a multidisciplinary perspective.

2.2 Autonomous Driving at a Glance

AVs, also referred to as self-driving vehicles, are intelligent vehicles equipped with advanced sensors, cameras, RADAR, LIDAR, GPS, and sophisticated learning algorithms that enable them to navigate and operate without human intervention [28]. To discern, identify, and distinguish the objects in their operational surroundings, these vehicles fuse information from a variety of sensors that help make real-time driving decisions [40, 290]. The history of contemporary AVs goes back to 1988, when ALVINN (Autonomous Land Vehicle In a Neural Network), the first neural network-powered self-driving vehicle taking camera images with a laser range finder, was able to produce control commands for the road-following task [213]. Current AVs deployed on road networks have different levels of automation based on their in-vehicle technologies and intelligent capabilities. SAE International has defined six levels of autonomous driving [245]: Level 0 - No automation (a human driver is responsible

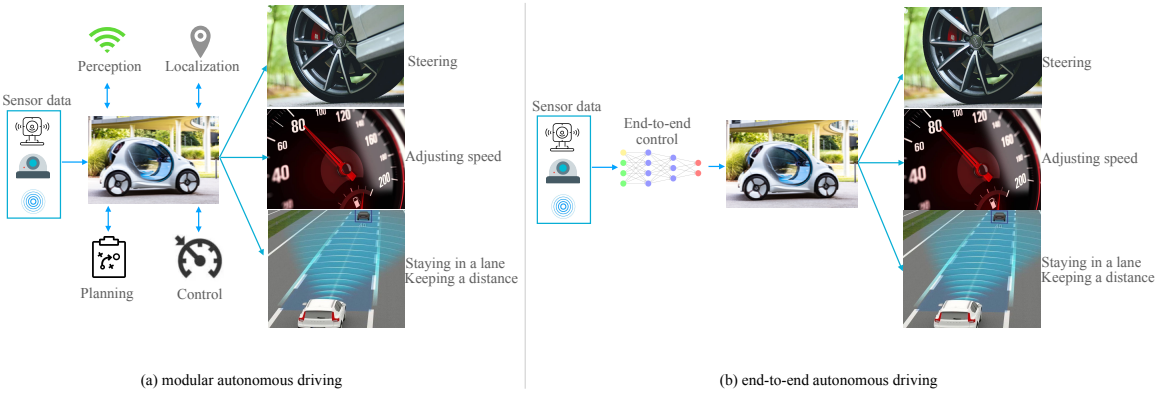


Figure 2.1: Modular vs. end-to-end autonomous driving. In the modular pipeline, the described operations are carried out subsequently to produce control commands, while end-to-end driving directly inputs raw sensor data and produces control commands as a unified task.

for all critical driving tasks); Level 1 - Driving assistance (a vehicle has automated driving support such as acceleration/braking or steering, but the driver is responsible for all other possible driving operations); Level 2 - Partial automation (Advanced Driving Assistance Systems (ADAS) operations such as steering and acceleration/braking are available at this level); Level 3 - Conditional automation (a vehicle has more advanced features such as object/obstacle detection and can carry out the majority of driving operations); Level 4 - High automation (a vehicle can fulfill all possible driving operations in a geofenced area); and Level 5 - Full automation (a vehicle can perform all driving operations in any likely scenario, and no human intervention is required) (see Table 2.1 for more details).

There are two main approaches to building autonomous driving systems in terms of their AI-based learning architecture: modular and end-to-end pipelines [43, 294]. The modular pipeline consists of four primary and interconnected modules, categorized as perception, localization, planning, and control (Figure 2.1, a). The modular pipeline leverages various sensor suites and algorithms for each module. While being comprised of standalone components makes the modular system more explainable and debuggable, such an architecture propagates errors to the next component, and thus, the overall pipeline error becomes cumulative [14, 43, 120].

Table 2.1: SAE International-defined AV levels with examples: While Levels 0-2 are human-supervised driving, Levels 3-5 are highly automated driving.

AVs level	**Description of the level**	**Example vehicles**
Level 0	A human driver is responsible for all critical driving tasks such as accelerating/braking, steering, etc. Driver support may include blindspot warning, automatic emergency braking, and lane departure warning.	Kia Rio 2004, Honda 2005, and other early and before 2000s' vehicles.
Level 1	The vehicle has automated driving support such as acceleration/braking or steering, but the driver is responsible for all other possible driving operations. Driver support may include lane centering or adaptive cruise control.	The majority of the daily used cars are in Level 1.
Level 2	ADAS operations, such as steering and acceleration/braking, are available. Nevertheless, the human driver should monitor the driving surroundings and take relevant actions when needed.	Toyota Corolla 2018, Nissan Sentra.
Level 3	The vehicle has more advanced features such as object/obstacle detection and can carry out most driving operations. However, human supervision is still required to take control of the car at any time.	The 2018 Audi A8 claimed to be the first commercial autonomous vehicle on Level 3 [226].
Level 4	The vehicle can fulfill all possible driving operations under specific conditions. The use of geofencing is required. Human supervision remains an option.	Alphabet's Waymo cars claim Level 4 automation [2].
Level 5	The vehicle can perform all driving operations in any likely scenario, and no human intervention is required. The Level 5 system can drive the vehicle in all possible road conditions.	There is no Level 5 vehicle in the current automotive industry.

In contrast to the modular pipeline, end-to-end autonomous driving has recently emerged as a paradigm shift in the design and development of AVs. End-to-end autonomous driving takes the raw sensor data as visual input and yields a control command for the vehicle (Figure 2.1, b) [43, 118, 257]. Particularly, recent breakthroughs in deep learning and computer vision algorithms, and the availability of rich sensor devices along with enhanced safety benefits have been the primary reasons for automotive researchers to leverage the end-to-end learning approach. The advantage of the end-to-end pipeline over its counterpart is that it directly produces driving actions by unifying perception, localization, planning, and control as one combined ML task. Furthermore, computational efficiency is improved via shared backbones in end-to-end learning, and in this way, potential information loss in intermediate layers is also avoided [120, 137].

2.3 Safety of Autonomous Driving

This section outlines key safety components and requirements of state-of-the-art autonomous driving technology. In this regard, the following subsections describe safety of autonomous driving systems from the aspects of software safety, hardware reliability, cybersecurity, human intervention in takeover situations, and fail-safe capability, respectively.

2.3.1 Software Safety

The safety of the deployed software system in autonomous driving is one of the most essential factors. As ML approaches play a pivotal role in enabling an automated vehicle to make intelligent decisions, these approaches need to be safety-compliant both theoretically and practically. The ISO 26262 standard has specifications to which the ML-based software architecture must conform [129]. Part 6 of this standard, known as V-model (Figure 2.2), is dedicated to product development at the software level for road vehicles, and state-of-the-art autonomous driving must comply with

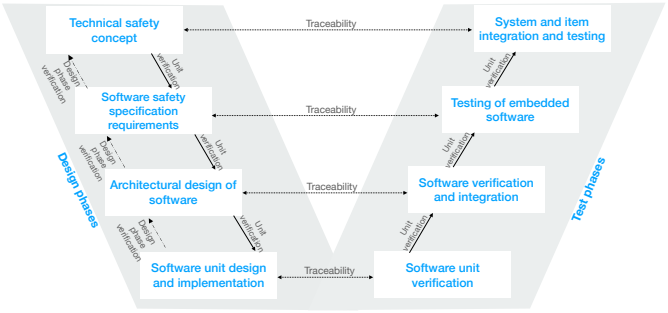


Figure 2.2: V-model in ISO 26262. The figure drawn based on the content in [129].

these stringent safety standards. The ISO 26262 standard has also established a risk classification system for the functional safety of automobiles, known as the Automotive Software Integrity Level (ASIL). This system groups risks into four categories, labeled A, B, C, and D [256]. ASIL A class indicates the lowest degree while ASIL D class indicates the highest degree of hazard. The complexity and associated risks with autonomous driving software deepen with respect to increasing SAE-defined automation levels, as a huge amount of computation happens on collected sensor data. According to Salay et al. [231], there are four safety issues with characteristics of ML on automotive software. The first one is the unstable behavior of a learning algorithm: As learning models are generally trained with local optimization algorithms, different training processes may yield different results on the same training dataset. Secondly, incompleteness of training is another negative factor, as only using some subset of input may not represent all safety-critical scenarios. Furthermore, despite the so-called "error rate" metric being an outcome of the ML models, whether this estimate is practically reliable remains unclear. Finally, the lack of interpretability of models or explainability of decisions is another issue for any ML-based intelligent driving system. The opaqueness of intelligent decision-making is a hindrance to safety assurance as it becomes harder for a user or examiner to trust if the model is operating as expected. Here, we analyze the latter point and analytically and empirically show that interpretable ML techniques have immense potential to improve safety of autonomous driving.

(a) Blind corner ahead (b) Disappearing lane markings and exiting

Figure 2.3: Potential takeover situations: (a) The blind corner ahead reduces an autonomous vehicle's perception ability, and (b) Autopilot perceives edge-markings of an exit lane as the current lane (upper left), steers right and the car exits the road incorrectly (upper right). The images have been adapted from [36].

2.3.2 Hardware Reliability

The reliability of autonomous driving systems depends on an appropriate combination of AI-powered software with hardware architecture. The physical architecture of a driving system should have seamless integration with its operating software in order to achieve reliable automated driving. ISO 26262 has defined functional safety of hardware systems by outlining the safety compliance process from the design specification to the production stage in modern autonomous cars [129]. These principles aim to identify and control hardware failure and prevent design faults in intelligent driving systems. These hardware failures could be *systematic* faults, arising during the design, development, and manufacturing stage, and *random* faults, occurring during the operational lifetime of an autonomous car due to possible casual defects, improper maintenance, and aging of hardware. Modern autonomous vehicles typically have the following major hardware components: sensors, actuators, V2X1 communication interfaces, and control hardware [143]. Sensors, such as vision cameras, lidar, radar, global positioning systems (GPS), and inertial measurement units (IMUs) enable an autonomous car to sense its operational environment dynamically and help

support real-time decisions based on the information fused from these data sources [290]. Actuators physically perform actions such as breaking, accelerating, and steering based on the obtained sensory information. Communication interface hardware powers a self-driving car to potentially interact with other surrounding vehicles, road users, and the infrastructure. Finally, control hardware ensures that a vehicle follows its predefined rule with proper steering angle, braking, and acceleration. Advanced driver-assistance systems (ADAS) are the most used control system in modern automated cars at present. So, as autonomous driving is a real-time decision-making process, these hardware components and their systematic integration must deliver low-latency, real-time, and high-speed data flow to the underlying ML software system.

2.3.3 Cybersecurity

The security of driving systems is yet another essential aspect of safe autonomous driving. ISO/SAE 21434 standard provides thorough guidelines and development strategies for the cybersecurity management of road vehicles [128]. With increasing reliance on big data, AI, and vehicular communication technologies, today's state-of-the-art autonomous vehicles are more prone to cyber-attacks. Such vulnerabilities, in turn, can impact driving safety adversely and cause severe damage to road users, such as passengers, pedestrians, and bystanders [212]. As two well-known examples of cybersecurity issues, Tencent Keen Security Lab controlled susceptible features of Tesla Models S, X, and BMW self-driving vehicles using remote attacks in their two subsequent studies [39, 140]. These studies have disclosed that an attacker could have easily controlled the mentioned self-driving cars both in parking and driving positions. Such flaws are a significant hazard to users' lives and demonstrate the need for more robust measures against cyber hacking.

2.3.4 Human Intervention in Takeover Situations

Autonomous cars may require takeover by backup drivers due to potential unexpected scenarios such as adverse weather conditions, approached construction zones, missed lane boundaries [183], and related issues in terms of software, hardware failures, and cybersecurity issues, covered above (see Figure 2.3). Vehicular takeover situations also become essential for highly automated driving at Level 3 and beyond. The transition of control from a car's intelligent driving system to the human operator happens in a short time interval and consists of two primary steps: 1) an autonomous car makes a takeover request (TOR), and a human driver should receive this request immediately and take over the control of the steering wheel and pedals, and 2) the post-takeover step, where the driver takes over the control of the car and manually performs the decisive action safely as per the traffic scenario [122]. The time granularity of these steps may vary from situation to situation, but the overall length of takeover cases is usually a few seconds. Consequently, a human actor must dominate takeover situations in the allotted time interval; otherwise, collision or other serious consequences may become inevitable. It is also necessary to specify that a self-driving car needs to be supplied with a relevant user interface or dashboard to alert the human driver on time and ensure a smooth transition from an autonomous mode to the human takeover mode.

2.3.5 Fail-Safe Capability

While the transition to highly automated driving is a significant change in the intelligent capabilities of self-driving vehicles, the immense need to operate safely in possible failure cases also escalates. A fail-safe is the ability of an automated car to take control and bring it to a safe stop position in case the primary autonomous functions fail or encounter an error and there is no backup driver in the vehicle [267]. The nature of failure may be hardware, software, or communication-related and in general, such breakdowns can be classified into two groups: AV component-based

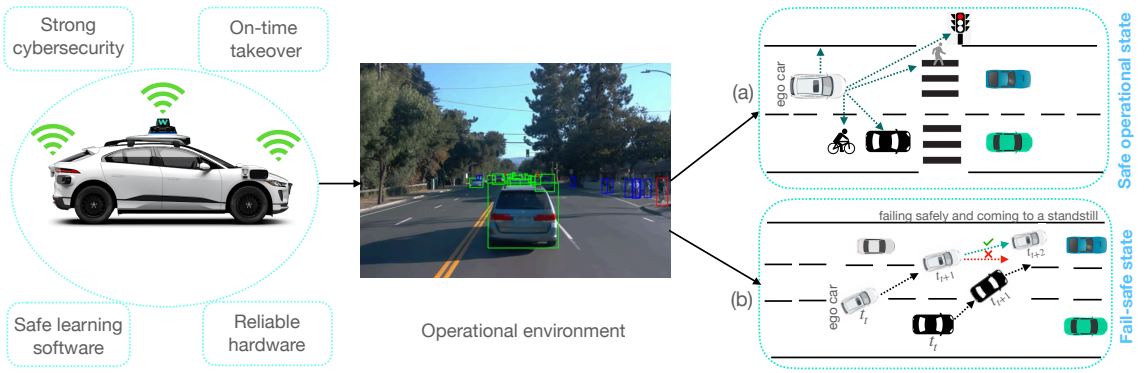


Figure 2.4: A diagram of safe autonomous driving. In (a), an autonomous car (i.e., ego car) interacts with the dynamic and stationary objects in the environment safely and keeps a distance from them. In (b), the ego car faces the unexpected action of the other vehicle, understands its limited motion ability at that moment, and comes to a standstill as it can not drive safely at that time step.

failures and infrastructure-based failures [29, 224]. The fail-safe ability in these cases can be achieved through a proper combination of sensor data, algorithms, and other vehicular technologies aiming to discern and respond to potential dangers while in motion. State-of-the-art solutions against potential AVs failures include road segment-based countermeasures (e.g., speed bumps, speed limit reduction pedestrian barriers), intersection-based countermeasures (e.g., high-visibility crosswalks, split phase timings), and effective infrastructure countermeasures (e.g., highways without potholes, well-maintained signs, and markings), in general [58]. In this sense, some studies have leveraged the concept of formal verification for fail-safe motion planning of self-driving vehicles [8, 208].

Overall, approaching an autonomous driving architecture as a holistic system, it is a considerably challenging task to achieve acceptable driving safety in real time. Reliable integration of cyber-physical, software components, and sensor suites is vital to accomplish safe autonomous driving and a fail-safe ability along the vehicle's motion trajectory (see Figure 2.4). Moreover, an AV must plan its trajectory considering the rational and unexpected behavior of human actors [220, 228] and predict future states of other dynamic objects [123] precisely to ensure safe driving in dynamic environ-

ments. In general, the established definition of safe autonomous driving requires risk minimization and assurance in seven key tasks namely, pedestrian detection, drowsy driver detection, vehicle detection, road detection, lane detection, traffic sign detection, and collision avoidance, as identified by [192], and an intelligent driving system must pass these safety checks while mapping sensor information to control commands.

2.4 Fundamental Issues

AI approaches, which are currently predominated by deep learning algorithms, have brought considerable improvements to many essential components of autonomous driving technology, including advances in perception, object detection, and planning. With the increasing level of automation, the number of AVs deployed to road networks has proliferated significantly in many developed European countries, the US, and Canada over the last decade [193]. However, the aforementioned road accidents involving such cars have caused public skepticism, and many studies have attempted to underscore the current limitations and issues with the design, development, and deployment of AV technology. For example, Fleetwood [91] has investigated public health and ethical issues arising with the use of autonomous driving. Their study provides an in-depth analysis of the health issues, especially with the Trolley problem examples [92, 136] (hitting a pedestrian on an icy road or a parked car; driving and hitting five people or changing the direction of the steering wheel and hitting an individual, etc.). Some studies have directly focused on the concept of ethical crashing (i.e., if crashing is inevitable, how to crash?) and the Trolley problem mentioned above. For instance, the Moral Machine experiment [24], a well-known and hotly debated experiment, investigates a general community's preferences on applied Trolley problems (inevitable accident scenarios with binary outcomes) and states that "these preferences can contribute to developing global, socially acceptable principles for machine ethics." However, further discussion on this issue condemns this opinion and draws attention to the lack of safety principles [174], which force deeper consideration

of such dilemmas [104]. Burton et al. [38] have identified three open problems in the state-of-the-art development of autonomous systems. The first one is the *semantic gap* that emerges when a thorough specification of the system is not provided to manufacturers and designers. Another identified issue is the *responsibility gap*, which arises when an accident happens, and the responsibility of either an autonomous system or a human actor as the cause of this accident remains unresolved. Finally, there is the question of who is responsible for compensating the injured during an accident, which precipitates the third issue: the *liability gap*. Burton et al.'s study [38] also reveals that the core of these issues is associated with domain complexity, system complexity, and transferring more decision-making functions from human actors to autonomous systems. Further studies include the outcomes of autonomous driving technology on public health in an urban area [247], and ethical dilemmas with AVs [180]. Overall, the key findings from these studies necessitate an understanding of the causes of these issues and intrinsically give the stakeholders the right to ask "why" questions.

2.5 Regulations and Standards

The issues and growing concerns caused by AI systems create the need to scrutinize the regulation of this technology. As a result, public institutions have initiated regulatory frameworks to monitor the activities of data-driven systems at both a country level and internationally. The focal points of these regulations are mainly to protect the stakeholders' rights and ensure they have control over their data. For example, the General Data Protection Regulation (GDPR) of the European Union (EU) initiated guidelines to promote the "right of an explanation" principle for users, enacted in 2016 and taking effect in May 2018 [221]. Moreover, the EU has a specially defined strategy on Guidelines of Trustworthy AI that has seven essential requirements, namely 1) human agency, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) accountability, 6) diversity, non-discrimination, and

fairness, and 7) societal and environmental well-being; these principles are all to be applied in AI-based product research and development [260].

Various organizations have recently proposed guidelines for the regulation of AVs to monitor their compliance with law enforcement. NACTO's (National Association of City Transportation Officials) statement on automated vehicles proposes nine principles to shape a policy on the regulation of future-generation AVs [194]. NHTSA of the US Department of Transportation has a specific federal guideline on automated vehicle policy to improve traffic safety [195]. In March 2022, NHTSA announced that automotive manufacturers would no longer have to equip fully autonomous cars with manual control elements, such as a steering wheel and braking pedals in the USA [68]. Canada [263], Germany [37], The UK [67], Australia [23], and Japan [4] have also recently launched their regulations on autonomous driving technology.

While the regulations have been set out to ensure legislative norms and user demands are met, some standards provide specifications to achieve a high safety level, quality assurance, efficiency, and environmentally friendly transportation systems. The International Organization for Standardization (ISO) has adopted several standards to define the relevant issues on automated driving. Examples include the ISO 21448 [126], which specifies situation awareness standards to maintain operational safety under the "Safety of the Intended Functionality," and the ISO 26262 [129] standard defined for the safety of electrical and electronic systems in production passenger vehicles, entitled as "Road vehicles – Functional safety." Detailed documentation of the legislation, regulation, and standardization of AVs can be viewed in [13].

2.6 Explanations in Autonomous Driving

2.6.1 The Need for Explanations

The need for explanations in autonomous driving arises from fundamental issues, established regulations and standards covered in previous subsections, and cross-

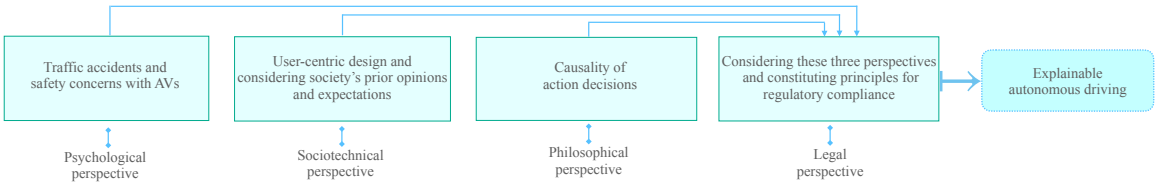


Figure 2.5: Cross-disciplinary factors necessitating explainability in autonomous driving

disciplinary views and opinions of society. At the highest level, the necessity of explanations for AVs can be summarized in terms of four perspectives:

- *Psychological* perspective: Traffic accidents and safety concerns remain the main cause of the need for XAI in autonomous driving from a psychological point of view [200].

- *Sociotechnical* perspective: The design, development, and deployment of AVs should be human-centered, reflecting the target audience's needs, and taking their prior opinions and expectations into account [71, 81].

- *Philosophical* perspective: Explaining AI decisions can provide descriptive information about the causal history of actions performed, particularly in critical situations [160, 188, 206].

- *Legal* perspective: It considers all the above-mentioned factors and incorporates them into general regulatory compliance principles for AVs. A notable example is GDPR's requirements on explanation provision for end users [221].

Overall, we can conclude that the explainability of autonomous driving systems is an expectation and a requirement from a multidisciplinary point of view.

2.6.2 Potential Benefits of Explanations for AVs

Considering these multi-dimensional perspectives, explainable autonomous driving can bring the following benefits to the stakeholders:

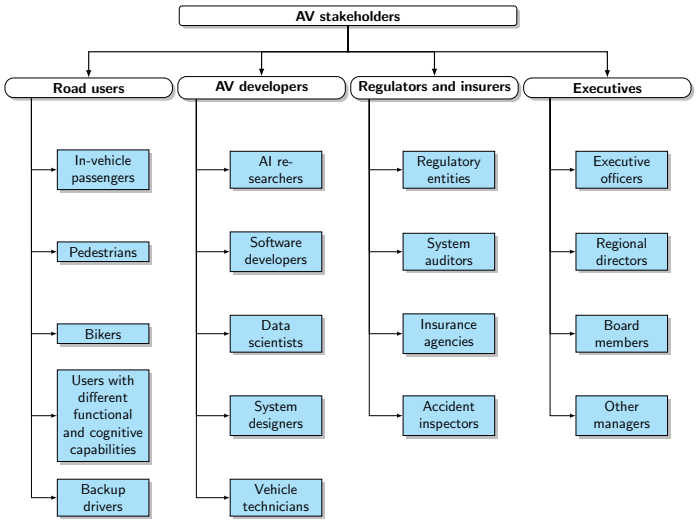


Figure 2.6: Taxonomy of the stakeholders in autonomous driving.

- *Human-Centered Design*: Getting the end users' inputs, opinions, and anticipations on the design and development of semi- or fully AVs can help with the acceptance of this technology by the general community [125].

- *Trustworthiness*: Algorithmic assurance can build trust in human-autonomous system relationships [130].

- *Traceability*: Explainable intelligent driving systems can help forensic analysts and system auditors understand the entire decision-making process of an AV via a post-trip analysis.

- *Transparency and accountability*: Explanations can help achieve accountability, which can resolve the potential liability and responsibility gaps in foreseeable post-accident investigations with the involvement of AVs as described by Burton et al. [38]. For example, Mercedes-Benz has recently taken a promising step forward and announced that the corporation will take legal responsibility for any accidents that their self-driving systems are engaged in [63]. Mercedes's declaration of legal culpability is a significant milestone toward the accountability of AV technology.

2.6.3 Explanation Recipients

The details, types, and delivery of explanations vary in accordance with users' identities, technical background knowledge in autonomous driving, and their various functional and cognitive abilities [15, 200]. For instance, a user having little technical expertise on how AVs operate may be satisfied with a simple explanation of a relevant decision/outcome. However, an autonomous systems engineer needs more informative explanations to understand the current functionalities of the car, with the motivation to appropriately "debug" the existing driving system as required. Therefore, the use of domain knowledge and expertise of the explainee is essential to provide pertinent, sufficiently informative, and intelligible explanations [157, 189]. Motivated by a target audience definition of [16] and [200], we can distinguish four groups of stakeholders from the perspective of explanation conveyance in autonomous driving, namely Group 1 - Road users, Group 2 - AV developers, Group 3 - Regulators and insurers, and Group 4 - Executive management of automotive companies. Figure 2.6 provides the identity of such stakeholders and their positions in the relevant classification.

2.6.4 Explanation Delivery Methods

As explainees are classified based on their domain knowledge and needs, explanations and their design and evaluation techniques also vary depending on the context and knowledge of the category of explainees. In fact, explanation construction is one of the major challenges in current XAI research. Zablocki et al. [295] define four "W" questions in XAI-based autonomous driving: 1) Who needs explanations? 2) Why are explanations needed? 3) What kind of explanations can be generated? and 4) When should explanations be delivered? In general, explanations in AI can be distinguished based on their derivation category and classification. Some of the early practical studies have applied explanations to automated collaborative filtering systems [111] and knowledge-intensive case-based reasoning [227]. Another empirical approach has attempted to derive explanations based on some intelligibility types

[164] and used "why," "why not," "what if," and "how to" type explanations for causality filtering. Furthermore, Liao et al. [162] have interviewed twenty user interface and design practitioners working in different areas of AI to understand users' explanatory requirements. By doing so, they have attempted to find the gaps in the interviewers' products and developed a question bank: the authors represent users' needs as questions so that users may potentially ask about the outcomes produced by an AI system. Overall, the stakeholder needs-based explanation design can be viewed as one of the promising approaches to effective delivery of explanation. Another popular approach to producing explanations is based on using psychological tools from formal theories, according to the literature review of [272]. Depending on the context and addressee, both explanation derivation methods confirm their usefulness. These explanation generation approaches can find alignment in their application in autonomous driving; since autonomous driving involves people with diverse backgrounds in society, relevant XAI design needs inherent adjustments to the context problem.

Except for informational content, the effective communication of explanations is also a key factor for good human-machine teaming [84]. In general, the conveyance of explanations to end users is realized through a user interface (UX) or a human-machine interface (HMI) [196]. For instance, an HMI may be an interface to alert the human driver to take over the control of a vehicle in an emergent situation. Other potential examples are heads-up displays, voice interfaces, light signals, and vibrotactile technology that explain the vehicle's decision-making intentions and bring situation awareness for people in the loop, as shown in [236].

Chapter 3

Explainable Artificial Intelligence Approaches for Autonomous Driving: A Comprehensive Overview

3.1 Introduction

This chapter presents a brief overview of frontier AI approaches - convolutional neural networks, recurrent neural networks, attention-based transformers, reinforcement learning, and imitation learning, before reviewing relevant XAI techniques. These AI techniques and their variations are dominant methods with empirical successes in the development of learning software for state-of-the-art AVs. This chapter is primarily based on the content in [20].

3.2 Artificial Intelligence for Autonomous Driving

3.2.1 Convolutional Neural Networks

Convolutional neural networks (CNN) are an AI architecture to process spatial information, such as images and videos [159]. As a powerful learning technique, CNNs can detect discriminant visual features automatically from an input image and are extensively used for pattern recognition, object classification and detection, and other computer vision applications. CNNs can be regarded as universal nonlinear function

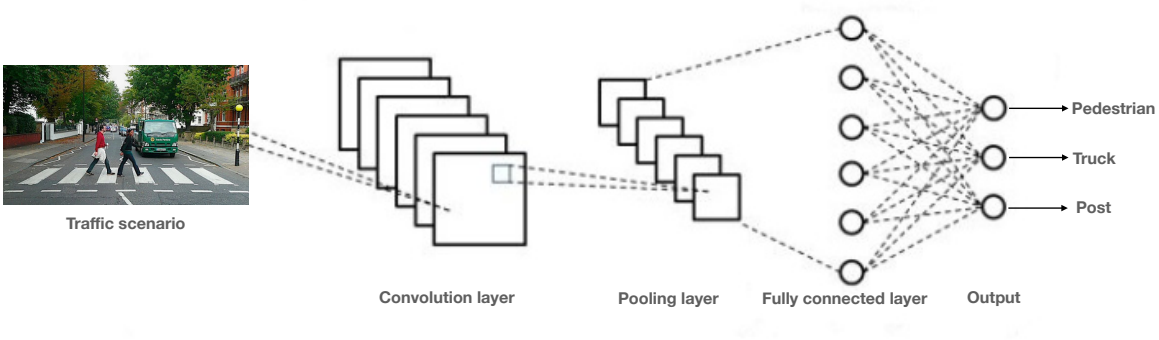


Figure 3.1: An example of a CNN for object classification in a real-time traffic scenario.

approximators. The input x of each layer in a CNN model consists of three dimensions: width, height, and depth. A typical convolutional neural network is parameterized by a weight vector consisting of a set of *weights*, W, between neurons and a set of *bias* values, b:

$$\theta = [W, b] \tag{3.1}$$

During training, a variety of useful and high-level features are extracted in the *convolution layer*. Then a *pooling layer* is used to reduce the size of the acquired feature map, to decrease computational costs. After that, the output of these steps is passed to the *fully connected layer* where neurons, along with the weights and biases, are connected with one another, and a nonlinear activation function is applied to the output of the previous step. The network further makes a final prediction. Commonly used activation functions with CNN are Sigmoid, Tanh, and ReLu, while ReLu is the most preferred because of its relatively faster convergence. In the context of autonomous driving, the role of CNN is indispensable for real-time scene understanding tasks, such as object detection, identification, segmentation, and classification. A typical example of CNN architecture for autonomous driving is shown in Figure 3.1.
While traditional CNNs have been successfully applied to numerous computer vision problems, a limitation of this learning architecture is that it can solely process

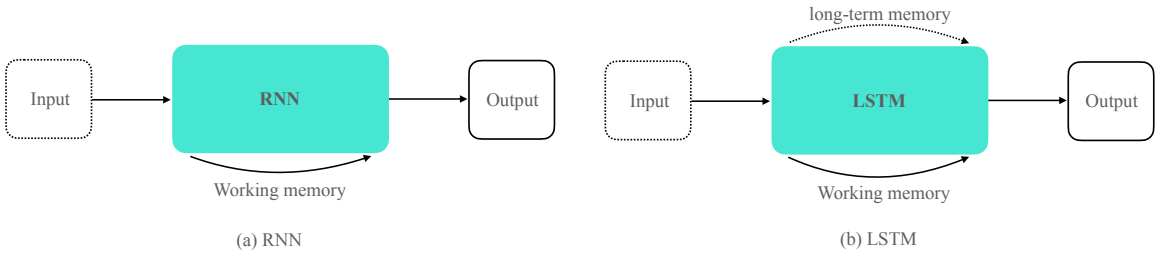


Figure 3.2: The network structure RNN and LSTM

Euclidean spatial data [35, 222]. However, in real-world applications, there might be a need to capture the spatial relationship between different elements necessitating the processing of non-Euclidean data. In this context, Graph CNNs (GCNNs) [150] have emerged as an alternative to conventional CNNs and are capable of processing any kind of non-grid data by representing elements in a graph structure. In autonomous driving, traffic data such as road networks, traffic flow, and spatial relationships among static and dynamic objects of such a network can be naturally represented as graphs, and GCNNs can effectively capture the dependencies and interactions among these elements. Moreover, as the driving environment is generally dynamic and rapidly changing, an inherent structure of GCNN can allow it to adapt to changes in the environment by dynamically updating the graph structure based on the provided sensor inputs and contextual information. Overall, both traditional CNNs and GCNNs and their augmented variations hold promise for dealing with a variety of challenging vision problems in autonomous driving.

3.2.2 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a deep learning architecture for processing temporal and sequential data, such as time series, video, and natural language data [108, 229]. RNNs have a feedback loop to iterate over time phases of sequential data: the output of a previous time step becomes an input to the current step. While iterating over the different time steps, recurrent networks can maintain internal states that contain information about each time step; thus, RNN architectures leverage the

concept of "memory," which uses information from a previous input to yield output in the next time step. A typical RNN architecture has three layers: input, hidden, and output layers. The input layer consists of N units. A sequence of vectors of each time step t denoted as $\{..., x_{t-1}, x_t, x_{t+1}, ...\}$ is the input of this layer. The input layer is connected to a hidden layer where connections between the units are defined by means of a weight matrix. The hidden units of a hidden layer connect with each other through recurrent connections, and by such a structure, the hidden layer defines the memory of the entire network, formulated as

$$h_t = f_H(o_t), \tag{3.2}$$

in which

$$o_t = W_{IH} x_t + W_{HH} h_{t-1} + b_h. \tag{3.3}$$

The hidden layer is also connected to the output layer with weights W_{HO}, and based on such a network flow, the units of the output layer are calculated as follows:

$$y_t = f_O(W_{HO} h_t + b_o). \tag{3.4}$$

Similar to CNNs, $f_O(\cdot)$ is the hidden layer activation function, and b_h is the bias vector of units of the hidden layer.

A major issue with traditional RNNs is the so-called *vanishing gradient problem* arising during training: network weights might not be effectively updated if the network is very deep. This may result in considerably small weight values that may reduce the network's learning ability and force RNNs to have a limited memory capacity. To solve this problem, long short-term memory (LSTM) has been proposed as an enhancement of RNN, which can handle sequential data more effectively and learn better [114]. Compared to a simple RNN architecture, LSTM has "gates" that control the flow of information through the network. Having the ability to preserve long-term

memory and retain earlier information in sequence makes LSTMs more practical than traditional RNNs. In general, both RNN and LSTM can be used to predict future position, velocity, and other parameters of autonomous driving. Figure 3.2 shows a structure of a typical RNN and LSTM.

3.2.3 Attention-based Transformers

Owing to their powerful representation capabilities, attention-based transformers have recently revolutionized the fields of natural language processing (NLP) and computer vision. While previous RNN-based encoder-decoder models powered the concept of attention for NLP applications, such as neural machine translation, the Transformer model introduces the self-attention mechanism that computes its input and output representations without employing RNNs or convolutions [265]. Given a sequence of items in various positions (e.g., words in a sentence), the representation of a sequence is calculated by finding relations or interactions between items in such a sequence. Let's denote a sequence of n items $(\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n)$ by $\mathbf{X} \in \mathbb{R}^{n \times d}$. Here, d is the embedding dimension representing each item. Self-attention aims to capture the interaction amongst all n items by encoding each item by using three learnable weight matrices to transform Queries ($\mathbf{W}^Q \in \mathbb{R}^{d \times d_q}$), Keys ($\mathbf{W}^K \in \mathbb{R}^{d \times d_k}$) and Values ($\mathbf{W}^V \in \mathbb{R}^{d \times d_v}$), in which $d_q = d_k$. Here, the input sequence $\mathbf{X}$ is first projected onto these weight matrices to obtain $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$ and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$. The final output $\mathbf{Z} \in \mathbb{R}^{n \times d_v}$ of the self-attention layer is then calculated as follows:

$$\mathbf{Z} = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}}\right)\mathbf{V} \tag{3.5}$$

For a given item, $\mathbf{x}_n$, in the sequence, the role of self-attention is to calculate the dot-product of the query with all keys, where the softmax operator is further used to get the attention scores. Each item, $\mathbf{x}_n$, eventually becomes the weighted sum of all items in the sequence $\mathbf{X}$, in which weights are given by the attention scores. The structure of the original attention-based Transformer is shown in Figure 3.3.

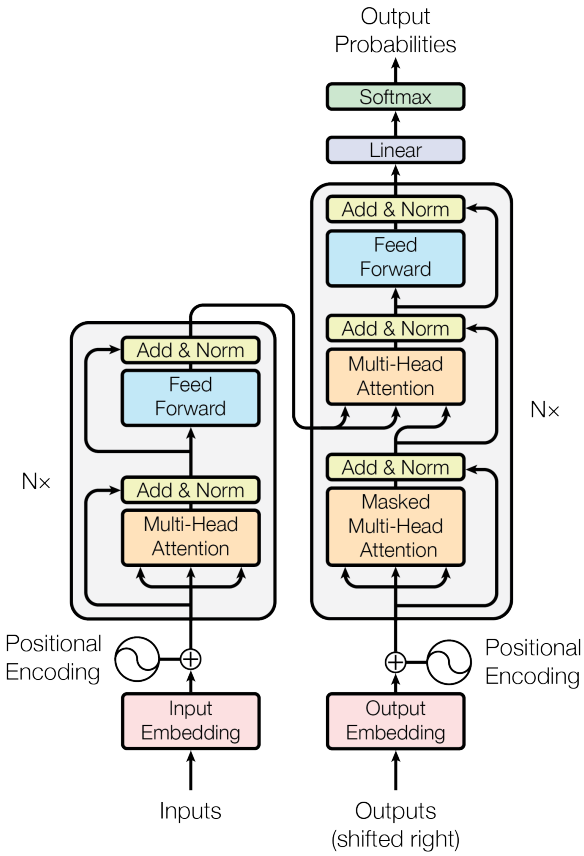


Figure 3.3: The network architecture of the original Transformer. Source: [265].

The impressive performance of the original attention-based Transformer model on machine translation has further found its successful applications in vision, forming a Vision Transformer [77]. The autonomous driving community has also further applied augmented variations of Vision Transformer to sensor fusion [49, 215, 242], semantic segmentation [45], point cloud recognition [155], and various object detection and recognition tasks [5, 293]. Thus, as AVs aim to map real-time vision to relevant action decisions, attention-based Transformer architectures have a huge potential to solve challenging vision and other relevant autonomous driving-related tasks.

3.2.4 Reinforcement Learning

Reinforcement learning (RL) is a learning approach where an autonomous software agent interacts with an operational environment and learns to improve its performance

via such interactions [255]. RL is a powerful machine learning framework for making real-time and sequential decisions. Generally, sequential decision-making problems are formalized within a setting formally known as Markov decision processes (MDP). An MDP comprises the following parameters:

- S - a set of states
- A - a set of actions
- T- a transition function
- R - a reward function
- γ - a discount factor defined as a fixed value in the (0, 1] interval.

In such a setting, selecting an action $a \in A$ results in a new state $s \in S$ with a transition probability $T(s, a, s') \in (0, 1)$, and gives a reward $R(s, a)$ to an agent. The goal of an RL agent (i.e., a self-driving vehicle, in our case) is to discover the optimal policy π^* that results in the maximum expected sum of discounted reward:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \underbrace{\mathbb{E}_\pi \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} \mid s_0 = s \right\}}_{:=V_\pi(s)}. \tag{3.6}$$

An agent's reward in starting from a state s taking action a by following a policy π is formulated as an *action-value function* (or Q-function) and defined as follows:

$$Q_\pi(s, a) = \mathbb{E}_\pi \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} \mid s_0 = s, a_0 = a \right\}. \tag{3.7}$$

Basically, the value function is a measure of how good it is for an agent to be in a particular state [255]. The horizon H is the number of time steps in a given MDP. The MDP setting for RL uses the Markov property: the current state transition depends only on the previous state and previous action. But often, in real-world problems, such as in autonomous driving, an agent might not be able to capture all information

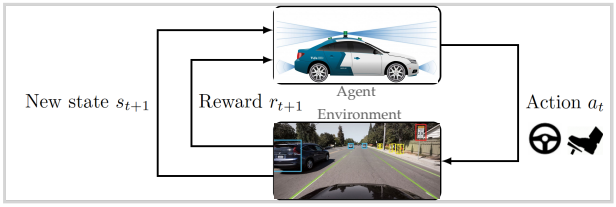


Figure 3.4: A diagram of reinforcement learning for autonomous driving.

and fail to observe all the states of the operational environment. In such a situation, an agent's interaction with the surroundings is constructed as a partially observable MDP (POMDP). In the POMDP setting, states are replaced by *observations* [254]. Observations are generated by a *latent state*, which is not available to an agent in a POMDP. The natural state for a POMDP setting is the distribution on the latent state; this is called a *belief state* [255]. We can infer that a typical observation is considerably less informative than a natural Markov state, S_t. So, depending on how perception defines the traffic scenario and tasks, both MDP and POMDP can be used for sequential decision-making in autonomous driving. A general diagram of RL for autonomous driving is described in Figure 3.4.

3.2.5 Imitation Learning

Imitation learning (IL) is a learning approach mimicking human behavior [124]. Also known as learning from demonstration, the agent learns from a dataset of demonstrations of human experts' behavior in contrast to an RL agent. The ultimate goal of IL is that the agent imitates the actions of a human expert, and its learned policy π_{il} matches the human expert's policy π_h. There are two well-known types of IL:

1) *Behaviour cloning (BC)*: In this category of IL, a model is trained in a supervised learning manner using a trajectory of state-action pairs $\{(s_0, a_0, s_1, a_1, \cdots)\}$. The goal is to minimize the loss function that calculates how the IL agent's actions are different from an expert's actions. The advantage of behavior cloning is that it is usually simple and efficient and does not require specific reward shaping. On the other hand, misleading correlations between input and output can affect the imitator's learning

ability seriously and lead to compounding errors.

2) *Inverse Reinforcement Learning (IRL)*: Also referred to as Inverse Optimal Control (IOC), IRL learns an underlying reward function that the human expert aims to maximize. Once the agent infers the expert's reward function, the optimal policy can then be learned via RL. Overall, empirical studies show that IRL can generalize better than BC in unseen states [240]; however, it is computationally costly and more intricate.

3.3 A Structure of the Literature Review

This section presents a comprehensive overview of XAI methods for AVs. In particular, I describe the classification of learning approaches in terms of vision, RL, IL, feature importance scores, logic, and user study-based explanations for AVs.

3.3.1 Visual Explanations

As deep neural networks, often in augmented forms of CNNs, power the vision ability of intelligent vehicles, understanding how CNNs capture real-time image segments that lead to the particular behavior of a vehicle is a key concept to achieving visual explanations. In this regard, explainable CNN architectures have resulted in adjustments to generate visual explanations. Zeiler and Fergus [296] use deconvolution layers to understand the internal representation of CNNs in their seminal work. Hendricks et al. [110] propose a model concentrating on distinguished properties of objects that explain the rationale for the predicted label. Zhou et al.'s [302] saliency map architecture, class activation map (CAM), highlights the discriminative part of an image to predict the label of the image. Moreover, Selvaraju et al. [239] propose an augmented version of CAM, called Grad-CAM, that highlights the derivative of CNN's prediction with respect to its input. Further examples of backpropagation-based methods include guided-backpropagation, [249], layer-wise relevance propaga-

tion [158, 232], and DeepLift [244]. Babiker and Goebel [26, 27] have also shown that heuristics-based Deep Visual Explanations (DVE) provide a justification for predictions of a CNN.

Explaining autonomous driving decisions using visual techniques is also primarily motivated by these studies. Particularly, Bojarski et al.'s work [32] is the first explainable vision approach for self-driving, where the authors propose a visualization method, called VisualBackProp, showing which set of *input pixels* contributes to a prediction made by CNNs. Their experiments conducted with the Udacity self-driving car dataset on an end-to-end autonomous driving task show that the proposed technique is a useful tool for debugging predictions of CNNs.

Similarly, Hofmarcher et al. [115] propose a *semantic segmentation model* implemented as a pixel-wise classification that explains underlying real-time perception of the environment. They evaluate the performance of their framework on Cityscapes [53], a benchmark dataset for understanding street scenes. The framework outperforms other popular segmentation models such as ENet and SegNet with 59.8 per-class mean intersection over union (IoU) and 84.3 per-category mean IoU. Interpretability of the model is a plus for unexpected behaviors, allowing to debug the driving system and understand the rationales for temporal decisions of a self-driving vehicle.

Kim and Canny [145] use a *causal attention* model on top of the saliency filtering that indicates which input regions actually affect the steering control. Their experiments are conducted on the driving datasets - Comma.ai [52], Udacity [264], and Hyundai Center of Excellence in Integrated Vehicle Safety Systems and Control (HCE): This model runs for nearly 16 hours to train CNNs end-to-end from images to steering angles and apply causality filtering to find out which parts of images have high influence in predictions (Figure 3.5). With this approach, the learned framework provides an interpretable visualization of a vehicle's actions. As an enhancement of this model, Kim et al. [147] provide textual explanations in their further study. They produce "intelligible explanations" on the action decisions of a self-driving vehicle using an

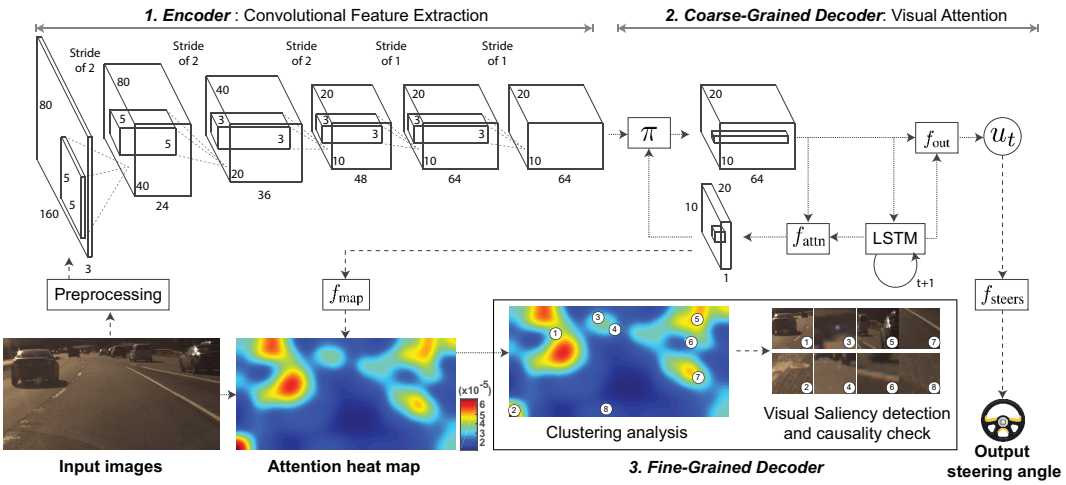


Figure 3.5: End-to-end learning of steering angle commands from an input image. Source: [145].

attention-based video-to-text mechanism and introduce a novel dataset called Berkeley DeepDrive-X (eXplanation) (BDD-X) that contains annotations for textual explanations and descriptions.

Zeng et al.'s [298] architecture learns to drive an autonomous vehicle safely by following traffic rules, including interaction with road users, yielding, and traffic signals. They use raw LIDAR data and an HD map to generate interpretable representations as 3D detection of objects, anticipated future trajectories, and cost map visualizations. 3D detection instances provide descriptive information so that the model understands the operational environment. Motion forecasting, measured as L1 and L2 distances, explains whether erroneous actions are due to incorrect velocity or calculation of direction. Finally, Cost Map visualization describes the traffic scene via a top-down view. The architecture is evaluated on a large real-driving dataset consisting of 6,500 traffic scenarios with 1.4 million frames and collected across several cities in North America, and measuring traffic rule violation, closeness to human trajectory, and collision. The authors also carry out an ablation study and show the impact of different overrides, input horizons, and training losses on end-to-end learning.

Xu et al. [285] propose *object-induced actions* with explanations for predictions of

Table 3.1: Studies on visual explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Bojarski et al., [32], 2016	Pixel-based explanations of CNN predictions	CNN	Visual	AV developers
Kim and Canny [145], 2017	Explaining behavior of a vehicle controller using heat maps	CNN, LSTM	Visual	AV developers
Kim et al., [147], 2018	Generating textual explanations on a vehicle's control commands	CNN, S2VT, LSTM	Visual and Textual	All groups
Hofmarcher et al., [115], 2019	Visual scene understanding using semantic segmentation	Enet, SqueezeNet 1.1, ELU	Visual	AV developers
Zeng et al., [298], 2019	End-to-end interpretable neural motion planner	FaF, IntentNet	Visual	AV developers
Hu et al., [119], 2019	Interpretable probabilistic prediction for autonomous driving	CVAE, Dynamic time warping, LSTM	Visual	AV developers
Xu et al., [285], 2020	Explaining object-induced action decisions for autonomous vehicles	Faster R-CNN	Visual	All groups
Kim et al., [146], 2020	Advisable learning by internalizing observation-to-action rules	Mask R-CNN, LSTM	Visual and Textual	All groups
Li et al., [161], 2021	Risk object identification via causal inference	InceptionResnet-V2, Mask R-CNN, Deep SORT, RoIAlign	Textual	All groups
Casas et al., [41], 2021	End-to-end model for mapless autonomous driving	CoordConv	Visual and Textual	All groups
Kim et al., [148], 2021	Explainable and advisable model for self-driving cars	DeepLab v3, Mask R-CNN, LSTM	Textual	All groups
Wang et al., [270], 2021	Enhancing automated driving with human foresight	Gaze-based vehicle reference	Visual	Road users
Chitta et al., [48], 2021	Interpretable neural attention fields for end-to-end driving	ResNet, MLP	Visual	AV developers
Dong et al., [74], 2021	Explainable autonomous driving via an image transformer	ResNet-50, Mobilenet-v2, multi-head self-attention	Textual	All groups
Hanna et al., [101], 2021	Interpretable goal recognition in the presence of occluded factors for autonomous vehicles	Goal and Occluded Factor Inference, Monte Carlo Tree Search	Visual	AV developers
Mankodiya et al., [177], 2021	XAI for trust management	Random Forest, Decision Tree, AdaBoost	Visual	AV developers
Madhav and Tyagi, [175], 2022	Explainable navigational intelligence for trustworthy autonomous driving	Grad-CAM, Lime	Visual	AV developers
Jing et al., [137], 2022	Interpretable action decision making for autonomous driving	Faster R-CNN	Visual and Textual	All groups
Jacob et al., [134], 2022	Region-targeted counterfactual explanations	GANs	Visual	AV developers
Zhang et al., [301], 2022	Interrelation modeling for explainable automated driving	Faster R-CNN, ResNet-50	Visual	AV developers
Kolekar et al., [152], 2022	Traffic scene understanding via U-Net and Grad-CAM	U-Net, GradCam	Visual	AV developers
Zemni et al., [297], 2023	Object-aware counterfactual explanations	BlobGAN	Visual	AV developers
Itkina and Kochenderfer [131], 2023	Trajectory prediction via interpretable self-aware neural networks	PostNet	Visual	AV developers
Feng et al., [88], 2023	Natural language explanations via semantic scene understanding	DeepLabV3	Visual and textual	All groups
Hu et al., [117], 2023	Interpretable trajectory prediction and decision-making of AVs	LaneGCNN, ResNet	Visual	AV developers
Dong et al., [73], 2023	Describing traffic scenes in natural language via attention-based transformer	CNN, LSTM, Transformer	Visual and textual	All groups
Atakishiyev et al., [17], 2023	Explaining autonomous driving actions with visual question answering	VGG-19, LSTM, DDPG	Textual	All groups
Echterhoff et al., [80], 2024	Leveraging concept bottlenecks as visual features for predicting control command and explanations of a car and human behavior	Longformer, GPT 3.5	Visual and textual	All groups
Feng and Sun [89], 2024	Interpreting self-driving decisions and improving safety by paying more attention to the regions that are near the ego vehicle	Multilayer Perceptron, Trajectory-guided Control Prediction	Visual	AV developers
Araluce et al., [14], 2024	Using driver attention for an end-to-end explainable decision-making from frontal images	ARAGAN, MobileNetV2	Visual	AV developers

an autonomous car. The authors introduce a new dataset called BDD-OIA, as an extension of the BDD100K dataset [291]; this extension is annotated with 21 explanation templates on a set of 4 actions. Their multi-task formulation for predicting actions also improves the accuracy of action selection. The CNN architecture further unifies reasoning on action-inducing objects and the context of scenes globally. The empirical results of the study on the introduced BDD-OIA dataset show that the explainability of the architecture also enhances action-inducing object recognition, resulting in better self-driving.

In two respective studies, Kim et al., [146, 148] propose an approach that leverages *human advice* to learn vehicle control (Figure 3.6). By sensing operational surroundings, the system is able to generate intelligible explanations of the action decisions (For example, "Slowing down *because* the road is wet"). The proposed architecture incorporates semantic segmentation with an attention mechanism that enriches knowledge representation. Experiments performed on the BDD-X dataset show that human advice with semantic segmentation and heat maps improves both the safety and explainability of predictive actions of a self-driving vehicle.

As a more recent vision-to-text approach, Atakishiyev et al. [17] employ the visual question answering (VQA) mechanism to explain autonomous driving actions. They train an RL agent and generate driving data showing the self-driving car's motion from its field of view. They further convert this video to image sequences, manually annotate the images with question-answer (QA) pairs, and encode questions and images with LSTM and pre-trained VGG-19 [246], respectively. The experimental results on five action categories show that VQA is a straightforward, effective, and human-interpretable approach to justify autonomous driving actions. Leveraging frontal images for interpretable decision-making has further been explored by subsequent studies as well [14, 73, 80].

While the mentioned studies focus on vision-based explanations of already obtained predictions of the model, there have been some recent studies paying attention to

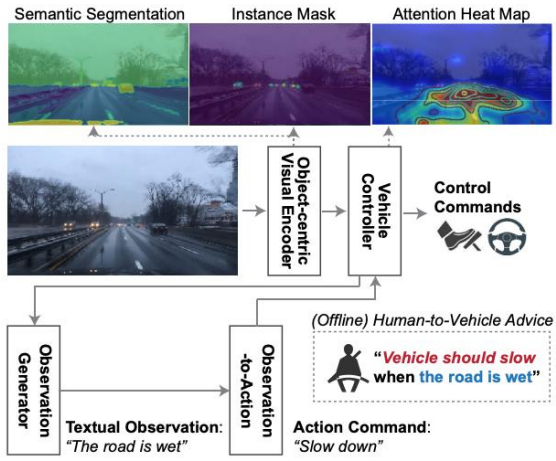


Figure 3.6: Human advice to a vehicle for appropriate action. Source: [146].

counterfactual explanations. In the context of automated driving, counterfactual analysis can be described with such an exemplary question: "Given the driving scene, how can it be modified so that the vehicle keeps driving instead of stopping ?" In other words, given the input, counterfactual analysis intends to figure out the distinguished features in this input that cause the model to make a certain prediction by envisioning modification of those features would cause the model to make a different prediction (e.g., Figure 3.7). Thus, in this case, the predictions obtained by the existing model and the imagined model become contrastive. As the application of counterfactual intervention, Li et al. [161] present an approach to find out risk objects that result in particular driving behavior. Their method, formalized as a Functional Causal Model (FCM), shows that the random elimination of some objects from the scene changes the driving decision to the contrastive prediction, such as from the "Stop" to "Go" command. In further work, Jacob et al. [134] introduce the STEEX model that uses a pre-trained generative model to produce counterfactual rationales by modifying the style of the scene while retaining the structure of the driving scene. Finally, as further enhancement of STEEX, Zemni et al. [297] propose a method called OCTET that generates object-aware counterfactual explanations without depending on the structural layout of the driving scene as backpropagation can optimize the spatial

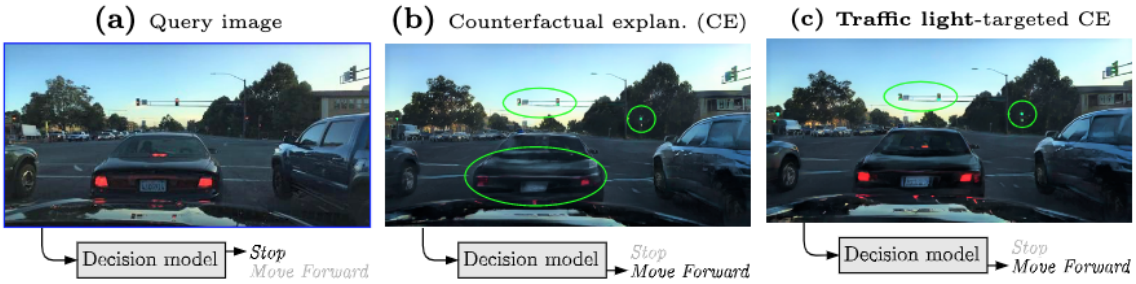


Figure 3.7: An example of a counterfactual explanation generated by STEEX. Graphics credit: [134].

positions of the provided instances.

Overall, we observe a significant focus and trend in visual explanations of autonomous driving systems, as such explanations provide an opportunity to better understand how accurately a self-driving vehicle senses the operational environment. Table 3.1 summarizes vision-based explanations for AVs.

3.3.2 Reinforcement Learning-based Explanations

While RL has had tremendous success in many sequential decision-making problems, including state-of-the-art autonomous driving tasks, explainable reinforcement learning (also known as explainable RL or XRL) is a considerably new, emergent research area and has not been investigated in a broad context [112, 268]. As self-driving is powered by real-time decision-making algorithms and explainability becomes an essential component of this field, explainable RL for autonomous driving becomes a promising topic of study. In this regard, we revisit propitious directions for explainable RL and provide their applicability for autonomous driving.

1) Explainable RL via Feature Importance: The feature importance (FI) concept in explainable RL is derived from classical supervised learning problems. While in supervised learning, the main input features that influence the model's prediction constitute FI scores, in the case of RL, this notion could be considered the state-action mapping [187]. LIME, SHAP values, and DeepLift are potential examples of such explanation provision techniques [268]. In the context of autonomous driving, FI

scores could be used to provide justifications for individual actions (such as turning left and accelerating) but are not generally capable of explaining the entire decisions of an intelligent driving system.

2) Explainable RL via Policy-level Explanations: Understanding an agent's long-term behavior is also a motivation for explainability in RL. In this regard, [187] proposes that policy-level explanations could be attained in three ways. Firstly, *transition summarization* could help understand which experiences have more influence on an agent's chosen action. In this way, one can interpret an agent's behavior while it explores. The second technique is to *convert a learned RL policy into an interpretable format*, often considering a transformation from an RNN to a finite-state machine (FSM). A third proposed approach is to extract *state clusters*. The idea is to understand an agent's actions by comparing selected actions in similar states. Moreover, Guo et al. [98] have shown that time step importance can also be used to obtain strategy-level explanations for an RL agent. Lastly, Kenny et al. [142] have recently proposed that leveraging human-friendly prototypes with a case-based reasoning technique (i.e., *I am taking action "a" because this state is similar to the previous prototypical state where I also performed "a.")* provides inherently produced explanations for a self-driving car's action.

3) Explainable RL via Reward-grounded Explanations: Another potential technique to interpret an agent's actions is to leverage a reward-based explanation, which is favored for intrinsically interpretable RL. One approach to reward-grounded explanation is *reward decomposition*. The main rationale behind this approach is to decompose the reward value into meaningful types in order to compare chosen actions in terms of such types [139]. Given C as a set of semantic reward types, the reward for a given state can be defined as:

$$R(s,a) = \sum_{c \in C} R_c(s,a) \tag{3.8}$$

Similarly, the action-value function can be defined as follows:

Table 3.2: Studies on RL-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Pan et al., [204], 2019	Semantic predictive control for explainable policy learning	LSTM, DDPG-SEG, DLA, model-based RL	Visual	AV developers
Chen et al., [42], 2021	Interpretable end-to-end autonomous driving with latent deep reinforcement learning	MaxEnt RL, DQN, DDPG, TD3 and SAC	Visual	AV developers
Schmidt et al., [235], 2021	Interpretable and verifiably RL for autonomoous driving learning	SafeVIPER, PPO	Visual	AV developers
Wang et al., [273], 2021	Learning interpretable end-to-end vision-based motion planning with optical flow distillation	IVMP, Optical flow	Visual	AV developers
Wang et al., [274], 2021	Uncovering interpretable internal states of merging tasks at highway on-ramps for autonomous driving decision-making	GMR, HMM	Visual	AV developers
Rjoub et al., [225], 2022	XAI-based federated deep RL for autonomous driving	DQN, DQN-XAI, SHAP	Visual	AV developers
Hejase et al., [109], 2022	Interpretable state representation for deep RL in autonomous driving	DDQN	Visual	AV developers
Paleja et al., [202], 2023	Interpretable continuous control trees for autonomous driving	Differentiable Decision Trees, SAC	Visual	AV developers
Kenny et al., [142], 2023	Interpretable Deep RL with Human-Friendly Prototypes for autonomous driving	PW-Net, PPO, TD3	Visual	AV developers
Yang et al., [288], 2023	Reward consistency for interpretable feature discovery for autonomous driving	PPO	Visual	AV developers
Lu et al., [171], 2024	Human-like cognitive maps for enhancing interpretability of autonomous driving	Successor Representations, Cognitive Potential Field	Visual	AV developers
Wang and Aouf [271], 2024	Explainable deep adversarial RL for robust autonomous driving	PPO	Visual	AV developers

$$Q^{\pi}(s,a) = \sum_{c \in C} Q_c^{\pi}(s,a) \tag{3.9}$$

With this motivation, one can understand how established types of rewards contribute to an agent's actions. The concept of reward decomposition has recently been used in several tasks, such as in Real-time Strategy games [10], a real-world medical dataset, MIMIC III [30], and Atari and MuJoCo games [98]. In another reward-driven interpretability technique, Yang et al. [288] have recently shown that reward consistency can be leveraged to solve the gradient disconnection on reward-action mapping. Based on these two proposals, we see that the reward mechanism by itself has good potential for achieving interpretable RL approaches for autonomous driving.

4) Logic-based Explanations: Logic-based explanations in the context of RL refer to the use of formal logic and a combination of neuro-symbolic approaches to interpret decisions made by an RL agent. This direction is a very recent trend in the XRL

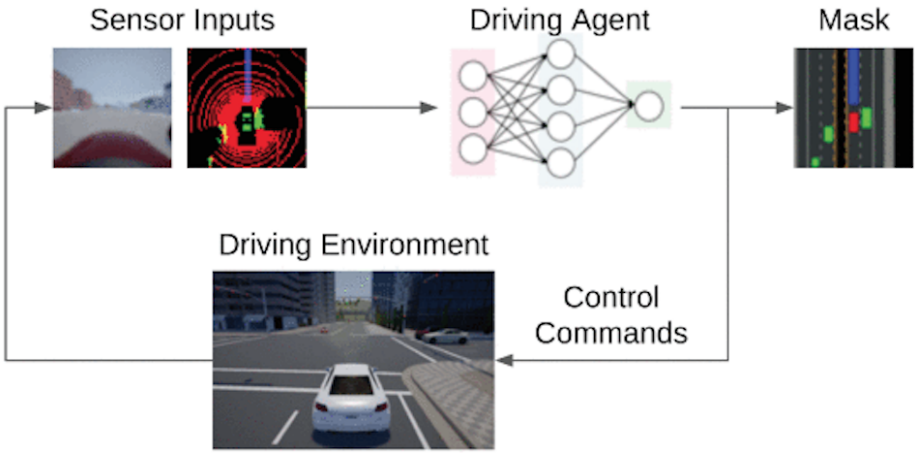


Figure 3.8: RL-based interpretable end-to-end autonomous driving via a bird-eye mask. Credit: [42]

literature. As far as I know, the only example of logic-based explanations in RL for autonomous driving is Song et al.'s work [248], where the authors leverage differentiable logic machine to learn an interpretable controller in lane changing and intersection management scenarios.

5) Explainable RL via Saliency Maps: Similar to other learning paradigms, visualization techniques can also make an RL process more interpretable. In this sense, using saliency maps, also known as heatmaps, is one of the potential approaches for explaining an RL agent's decisions. Some well-known examples include the visualization of Atari agents [97], where the authors visualize agents to understand the evolvement of their underlying policy. Interestingly, the further work [22] suggests that saliency maps can be thought of as an exploratory tool rather than an interpretability tool in the case of Atari agents. Finally, as the last study of investigating Atari agents with heatmaps, "what if" type counterfactual state explanations have been proposed [198]. In general, saliency maps mostly provide subjective explanations in RL, and more thorough studies are needed to draw strong explainability conclusions with this technique [187, 268].

6) Explainable RL via Natural Language Explanations: Finally, another promising approach to generating explanations on RL decisions is linguistic explanations con-

veyed via a natural language. As covered in Section III.B, end users in autonomous driving have different backgrounds and varying degrees of technical knowledge, so explanation construction should take this factor into account. In this regard, natural language explanations are viewed as the best form of explanation as this category of explanations is simple, informative, intelligible, and can be understood by all consumers. With that said, in the context of autonomous driving, explaining RL policies/actions in natural language has not been well-studied in the current literature. Within the vision context, one of the prominent works is Kim et al.'s work, where authors generate text-based causal explanations via an attention-based video-to-text model [147]. Some further studies include providing descriptive information about actions using natural language [73, 88, 134, 285]; however, these explanations are mapped from visual inputs. As RL approaches are becoming more applicable to highly autonomous driving, there is an immense need to generate human-understandable explanations for the RL-based decision-making of AVs. Natural language explanations of an RL agent's actions have recently been studied for general sequential decision-making problems [210]; however, to the best of our knowledge, the concept has not been explored empirically yet in the context of autonomous driving. A summary of RL-based explanations for AVs is provided in Table 3.2.

3.3.3 Imitation Learning-based Explanations

Similar to explaining RL-based autonomous driving, there have also been few studies to explain IL-based agents in the field via techniques similar to RL-based explanations. Cultrera et al. 's work [60] is one of the preliminary studies in this context. The study uses a conditional IL agent along with the attention model in the CARLA simulator [76] to produce an explanation, describing the influence of distinctive parts of the driving images on the predictions of the model.

Hu et al. [116] further introduce model-based IL (MILE) to learn a latent space in the driving environment via high-resolution videos from demonstrations of an expert

Table 3.3: Studies on IL-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Cultrera et al., [60], 2020	Explaining autonomous driving by learning end-to-end visual attention	CNN, IL	Visual	AV developers
Teng et al., [259], 2022	Interpretable imitation learning for end-to-end autonomous driving	Bird's Eye View model, IL	Visual	AV developers
Renz et al., [223], 2022	Explainable planning for autonomous driving	BERT, GRU, IL	Visual	AV developers
Hu et al., [116], 2022	Model-based IL for autonomous driving	IL	Visual	AV developers
Liu et al., [170], 2024	Interpretable generative adversarial IL for autonomous driving	IL, Signal Temporal Logic	Visual	AV developers

in a simulation environment. The experiment shows that the proposed model can predict feasible states and actions for an IL agent and present interpretability of driving decisions via a bird's-eye-based semantic segmentation map.

Another bird's eye view-based interpretation technique within IL for end-to-end driving has been developed by Teng et al. [259], as a two-stage model called Hierarchical Interpretable Imitation Learning (HIIL). In the first stage, a bird's eye view is presented to interpret the environment. After that, the latent features from the first stage are acquired along with the steering angle from the Pure-Pursuit algorithm [55]. In this manner, an agent's actions are predicted in a simulation environment, and the experimental results show that HIIL demonstrates enhanced interpretability along with generalization in complex driving scenarios.

Moreover, Renz et al. [223] propose a transformer architecture, PlanT, based on IL with an object-level representation of the input. By incorporating driving information about other vehicles in the driving environment, and 360° field-of-view of a self-driving vehicle, PlanT achieves superior performance and explains the self-driving car's decision-making rationale with visual cues in the simulation environment.

Finally, in the most recent work, Liu et al. [170] introduce a generative adversarial network (GAN)-based IL approach combined with signal temporal logic (STL) and test the applicability of the model in a simulator. The experiment collected in the CARLA simulator shows that their IL model combined with STL is highly effective

Table 3.4: Studies on feature importance-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Omeiza et al., [199], 2021	Generating tree-based explanations with and without causal attributions	Tree-based representation / User study	Textual	All groups
Brewitt et al., [33], 2021	Interpretable and verifiable goal recognition with learned decision trees for autonomous driving	Decision Tree	Visual and Textual	AV developers
Mankondiya et al., [177], 2021	XAI for trust management in autonomous vehicles	Random Forest, Decision Tree, AdaBoost	Visual	AV developers
Cui et al., [59], 2022	Interpretation framework for autonomous driving	Random Forest, SHAP	Visual	AV developers
Onyekpe et al., [201], 2022	AV positioning using SHAP	SHAP, WhONet	Visual and Textual	AV developers
Almalioglu et al., [7], 2022	Vehicle position with deep learning	GRAMME, SHAP	Visual	AV developers
Ayoub et al., [25], 2022	Predicting driver takeover time in conditional automated driving	XGBoost, SHAP	Visual	AV developers
Brewitt et al., [34], 2023	Interpretatable trees for goal recognition in autonomous driving under occlusion	OGRIT, Decision Tree	Visual	AV developers

in adjusting rules to adapt the driving policy for unseen situations. Table 3.3 summarizes the reviewed studies on IL-based explanations for AVs.

3.3.4 Feature Importance-based Explanations

Being inherently interpretable and easier to understand a prediction of a model, feature importance scores have also been investigated for interpretability purposes in autonomous driving. These studies primarily use decision tree-based metrics and SHAP values [173] to quantify the contribution of each feature in predictions of a driving model, thus providing a holistic understanding of how inputs influence an autonomous car's real-time actions. Decision trees have been proven to describe the rationale semantically for each prediction made by a CNN architecture [300]. Omeiza et al. [199] use decision trees as a *tree-based* representation that generates scenario-based explanations of different types by mapping observations to actions in accordance with traffic rules. They use human evaluation in a variety of driving scenarios and generate Why, Why Not, What If, and What explanations for driving situations and empirically prove that the approach is effective for the intelligibility and accountability goals of automated vehicles.

Brewitt et al. [33] introduce Goal Recognition with Interpretable Trees (GRIT) and

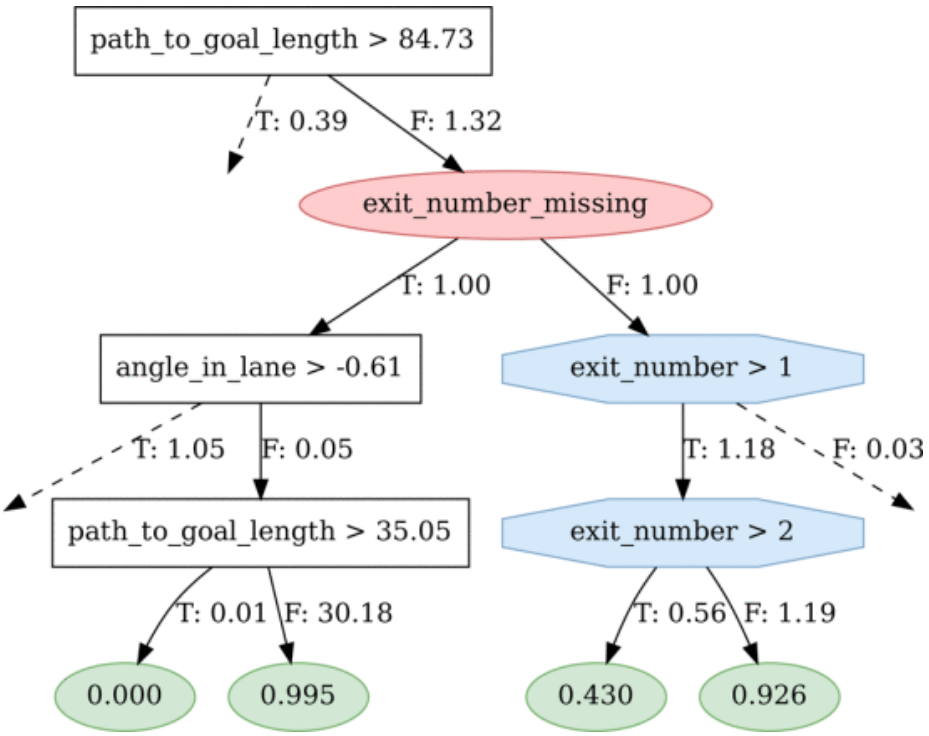


Figure 3.9: An interpretable decision tree for an exit-roundabout scenario as verifiable goal recognition. Credit: [34]

its augmented version, OGRIT [34], leveraging decision trees trained from the trajectory data of a self-driving car. Figure 3.9 from [34] shows an example of interpretable decision tree for exit-roundabout goal type. The frameworks, tested on various driving scenarios are proven empirically verifiable for goal recognition using a satisfiability modulo theories (SMT) solver [65].

Cui et al. [59] use Random Forest for the interpretability purpose on the autonomous car-following task. They employ deep RL for the decision-making of an autonomous car and employ SHAP values to simplify the feature space. Once the agent generates state-action pairs, Random Forest is applied to these pairs and experimental results show the approach works effectively to explain behavior for the designated car-following task. In a recent study, Random Forest has also been proven to detect misbehaving vehicles in Vehicular Adhoc Networks (VANET) in Mankodiya et al.'s work [177]. Thus, being computationally more transparent than traditional deep

neural network architectures, decision trees and SHAP can explain a variety of autonomous driving tasks with less computation. Table 3.4 summarizes the reviewed studies on feature importance-based explanations for AVs.

3.3.5 Logic-based Explanations

While the interpretability of a deployed autonomous driving control model has been the dominant direction for research, there have also been attempts to verify the safety of self-driving vehicles with logical reasoning. In this regard, Corso and Kochenderfer [54] present a technique to identify interpretable failures of autonomous cars. They use *signal temporal logic* expressions to describe failure cases of an autonomous car in an unprotected left turn and pedestrian crossing scenarios. For this purpose, the authors use genetic programming to optimize signal temporal logic expressions that acquire disturbances trajectories, causing a vehicle to fail in its decisive action. The experimental results show that the proposed approach is effective for interpreting safety validation of an autonomous car.

Suchan et al. [251] have developed an *answer set programming*-based abductive reasoning framework for online sensemaking for perception and control tasks. In its essence, the framework integrates knowledge representation and computer vision in an online manner to explain the dynamics of traffic scenes, particularly occlusion scenarios. The authors demonstrate their method's explainability and commonsensical value with empirical study collected on the KITTI MOD [96] dataset and the MOT benchmark [186]. Another experimental study leveraging the concept of answer set programming has been carried out by Kothawade et al. [154]: they introduce AUTO-DISCERN, a system that incorporates common sense reasoning with answer set programming to automate explainable decision-making for self-driving vehicles. They test their rules and show AUTO-DISCERN's credibility in real-world scenarios, such as lane changing and right turn operations, from the KITTI dataset. Table 3.5 summarizes the reviewed studies on logic-based explanations for AVs.

Table 3.5: Studies on logic-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Suchan et al., [251], 2019	An answer set programming-based abductive reasoning for visual sensemaking	Answer set programming, YOLOv3, SSD, Faster R-CNN	Visual	AV developers
Corso and Kochenderfer [54], 2020	Interpretable safety validation for autonomous driving	Signal temporal logic	Textual	AV developers
DeCastro et al., [66], 2020	Interpreting policies via signal temporal logic for autonomous driving	Signal temporal logic, LSTM, CVAE	Visual	AV developers
Kothawade et al, [154], 2021	Explainable autonomous driving using commonsense reasoning	ASP, s(CASP)	Textual	Road users

3.3.6 User Study-based Explanations

Some investigations involve users in case studies to understand the effective strategies for explanation generation in autonomous driving tasks. The key idea of a user study is that getting people's input in designated driving tasks can help improve the adequacy and quality of explanations in autonomous driving. Wiegand et al. [282] perform a user study that identifies a mental model of users for determining an effective practical implementation of an explanation interface. The main research question here is to understand what components need to be visualized in a vehicle so the user can comprehend the decisions of self-driving vehicles. The study discloses that combining an expert mental model with a user mental model as a target mental model enhances the drivers' situation awareness. Furthermore, Wiegand et al. [281] investigate situations in which explanations are needed and methods relevant to these situations. They spot seventeen scenarios where a self-driving vehicle behaves unexpectedly. Twenty-six participants are selected to validate these situations in the CarMaker driving simulator to provide insights into drivers' need for explanations. As a result of the user study, the authors identify six groups to highlight the primary concerns of drivers with these unexpected behaviors, namely emotion and evaluation, interpretation and reason, the capability of a self-driving car, interaction, driving forecasting, and request times for explanations.

Wang et al. [270] propose an approach that enables a human driver to provide *scene forecasting* to an intelligent driving system using a purposeful gaze. They develop a

Table 3.6: User study-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Wiegand et al., [282], 2019	Explaining driving behavior of autonomous cars	User study	Textual	Backup drivers
Wiegand et al., [281], 2020	Understanding situations that a driver needs explanations	User study	Visual	All groups
Wang et al., [270], 2021	Enhancing automated driving with human foresight	User study	Visual	Backup drivers
Schneider et al., [237], 2021	UX for transparency in autonomous driving	UEQ-S, AVAM (User study)	Visual, Textual, Light	All groups
Schneider et al., [236], 2021	Increasing UX through different feedback modalities	UEQ-S (User study)	Visual, Textual, Audio, Light, Vibration	All groups
Shen et al., [243], 2022	Identifying which scenarios need explanations in autonomous driving	Friedman test, Pearson correlation, Point-Biserial Correlation	Visual	Road users
Schneider et al., [238], 2023	The role explanatory information in failure situations in highly autonomous driving	UEQ-S, AVAM	Visual, Textual	All groups
Kim et al., [144], 2023	Timing perspective and mode of explanations for road users in autonomous driving	GradCam, Head-mounted display, Windshield display	Visual	Road users

graphical user interface to understand the effect of human drivers on the prediction and control of an intelligent vehicle. A simulator is used to test and verify three driving situations where a human driver's input can improve safety of self-driving.

Apart from these works, Schneider et al. involve human participants in their empirical studies to understand the role of explanations for the public acceptance of AVs [236, 237]. They explore the role of explainability-supplied UX in AVs, provide driving-related explanations to end users with different methods, such as textual, visual, and lighting techniques, and conclude that providing context-aware explanations on autonomous driving actions increases users' trust in this technology. Their subsequent study also confirms that driving explanations can help mitigate the negative impact of AVs failures on users [238]. Finally, Kim et al.'s user study [144] confirms that humans do not need explanations seamlessly, and presenting explanations only in critical driving conditions is preferred to enjoy the trip with an autonomous car and prevent information overload. Table 3.6 summarizes the reviewed user studies on XAI for AVs.

A high-level overview of all these studies indicates driving explanations are generally multi-modal, context-dependent, and task-specific. Moreover, end-to-end learning is

gradually becoming more popular for highly autonomous decision-making owing to powerful deep-learning approaches and overall safety and efficiency benefits. Based on the insights from the state of the art, we can define explainable autonomous driving as follows:

> *Explainable autonomous driving is a self-driving approach powered by a compendium of AI techniques 1) ensuring an acceptable level of safety for a vehicle's real-time decisions, 2) providing explanatory information on the action decisions in critical traffic scenarios in a timely manner, and 3) obeying all traffic rules established by the legal entities and regulators.*

Driven by this definition and the state-of-the-art works above, in the subsequent section I present a general and conceptual XAI framework for end-to-end autonomous driving aligned with industrial trends and show the necessary components and steps to achieve regulatory-compliant AVs in the next generation.

3.4 A Conceptual Framework for Explainable Autonomous Driving

I propose a general framework in which methods for developing XAI, end-to-end learning, and safety components are combined to inform processes of regulatory principles. Each of these components has a concrete role in our framework. In my recent study [21], I have briefly described end-to-end learning for AVs. I extend the scope of that work and describe the essential elements of end-to-end autonomous driving, and the role of and potential challenges with explanations in such a setting. These individual components are described as follows:

1. An end-to-end control component: Given all possible instances of environment,

$$E = \{e_1, e_2, ...e_n\},$$

and a compendium of actions

$$A = \{a_1, a_2, ...a_n\},$$

an autonomous car can take, the overall role of a *control system* is to map the perceived environment to corresponding actions:

$$C : E \mapsto A.$$

This mapping intends to ensure that a controller maps the environment to a relevant action of an autonomous system. A control system C is an *end-to-end control system* (***eeC***), if C is a total function that maps every instance of an environment

$$e \in E$$

to a relevant action

$$a \in A.$$

The most prevalent learning paradigms for end-to-end autonomous driving are RL and IL [43]. Furthermore, differentiable learning has also recently emerged as an end-to-end driving architecture: While the planning component is prioritized, this learning pipeline optimizes several modules of the entire driving architecture (e.g., [120]). Overall, as described in Chapter 2, the end-to-end learning pipeline uses a single deep neural network as a unified task to map the sensor model of the world to real-time control commands of AVs.

2. A safety-regulatory compliance component: The role of the safety-regulatory compliance component, ***srC***, is to represent the function of a regulatory agency, one of whose main roles is to verify the safety of any combination of eeC with autonomous vehicle actions A:

$$srC = f(eeC, A).$$

This requirement could be as pragmatic as some inspection of individual vehicle safety (for example, verifying basic safety functions of an individual vehicle for re-

licensing). That said, this concept should be considered as a thorough compliance testing of *eeC* components from vehicle manufacturers to certify their public safety under international and/or national transportation guidelines such as [95] and [263]. The general principles for acceptable functional safety of road vehicles are defined by the ISO 26262 standard [129]. According to this standard, there should be a safety certification development with evidence-based rationales: the vehicle should be able to meet the established functional safety requirement in its operational context. Part 6 of the ISO 26262 standard [127] is dedicated to end-product development for automotive applications within the software level. This guideline includes the design, development, testing, and verification of software systems in automotive applications. Based on these standards, there seem to be two fundamental approaches to confirming regulatory compliance, which we label confirmation of compliance by "simulation," and confirmation of compliance by "verification." These steps are aligned with my observation regarding the role of XAI in confirming regulatory compliance. In the case of the process of establishing regulatory compliance by *simulation*, the idea is that a selected set of autonomous actions can be simulated, and then assessed to be satisfactory. This approach is perhaps the most familiar, as it arises naturally from an engineering development trajectory, where the accuracy of simulators determines the quality of compliance (e.g., [138]). The confidence of the established compliance is a function of the accuracy and coverage of the simulation. However, this compliance process can be very expensive and prone to safety gaps, especially when consensus on the properties and scope of a simulation is difficult to achieve. Thus, in general, the simulation part can be considered a "driving school" for AVs: The designed and developed learning software system should be tested rigorously in this phase before such an autonomous system, as a holistic architecture, is deployed to a real vehicle in the physical environment and real roads. The alternative, *verification*, is aligned with the proposed framework and has significant foundational components established in the discipline of proving software correctness, with a long history (e.g., [103]). The

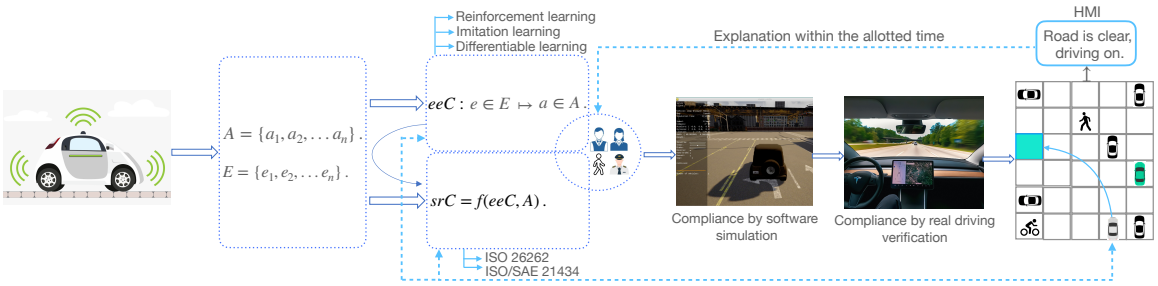


Figure 3.10: A diagram of the proposed explainable end-to-end autonomous driving framework.

general idea is that offline simulation-based autonomous driving is validated on real roads on real AVs via real sensor suites and a learning software stack by passing the safety checks of regulatory compliance.

In addition to safety assurance, another critical requirement of AVs is their ability to defend against adversarial attacks. The ISO/SAE 21434 standard has defined guidelines for cybersecurity risk management for road vehicles, and AVs must also comply with these requirements [128]. As AVs increasingly rely on their automation ability, it is vital that ML software of an intelligent driving system and built-in interfaces can detect and defend against potential cyber-attacks of the broad spectrum, such as electronic control units (ECU) attacks, in-vehicle network attacks, and automotive key-related attacks [149, 217, 253].

We can expect that the potential evolution of the srC processes will ultimately rely on the automation of regulatory compliance testing against all eeC systems. The complexity of srC systems lies within the scope of the testing methods established in a legal framework, where these methods are the basis for confirming a threshold of safety. For instance, a regulatory agency may require at least 90% regulatory-compliant performance of any particular eeC from N safety tests to be performed. However, as a general requirement, this performance must meet ISO 26262 and ISO/SAE 21434 standards to ensure that an autonomous car's decision-making procedure is aligned with its underlying ML software: The safety features must pass critical checkpoints, and the autonomous car has to have the ability to defend itself against foreseeable

adversarial attacks.

3. An explanation component: This constituent of the framework provides understandable insights into real-time action decisions made by autonomous driving, complying with and corresponding to an eeC and a srC. The explanation component must justify how the autonomous car chooses actions along the trip and has to be able to communicate these pieces of information to the relevant users both during the journey and via a post-trip analysis. As analyzed in the reviewed studies, explanations can be described in visual, textual, feature importance format or in hybrid, multi-modal ways and conveyed via light, audio, vibrotactile, and in other forms depending on the explainees' needs and preferences.

Temporal granularity and conveyance of explanations: While the format and content of explanations have been the primary focus of XAI research, it is noteworthy to underscore that another important consideration, the time granularity of explanations, has not been well-studied in the state of the art. In general, the timing perspective of AV explanations can be analyzed within three questions: 1) Should explanations be delivered before action is chosen or after action is performed? 2) What is the appropriate lead time for a safe transition from an automatic mode to a human takeover? and 3) Should explanations be delivered seamlessly or only when it is required? We analyze these nuances separately as follows:

1) *Time to deliver explanations*: Delivering timely explanations can help human drivers/passengers react to emergent situations, such as takeover requests, appropriately and prevent a potential danger in the vicinity. According to Koo et al.'s study [153], it is favorable to convey explanations before a driving event is about to happen. This concept has further been validated by Haspiel et al.'s user study, and human judgment shows that explanations should be delivered *before* action is *decided* rather than *after* it is *performed* [106]. This judgment makes sense as on-time communication of explanations can bring situation awareness for people on board and enable them to monitor an autonomous car's subsequent action. If the action to be per-

formed soon is hazardous, a human driver or passenger can manually intervene in the situation with such explanations and prevent potential danger ahead.

2) *The impact of lead time on the safe transition from an automated mode to a human takeover*: Another important criterion is determining the amount of time needed to alert human actors for a takeover request. In the user study measuring the impact of 4 s vs. 7 s as the lead time on takeover alert, Huang and Pitts [121] show that a shorter lead time leads to a faster transition to human-controlled mode but also lacks the quality of takeover as lack of time may be stressful for a human actor in such situations. A similar conclusion has been acquired by Mok et al. [191] in the case of 2, 5, and 8 s transition times. Wan and Xu [269] have further verified that an insufficient amount of lead time, such as 3 s, results in an impaired takeover performance, and drivers perform better when enough time, such as more than 10 s, is allotted for takeover requests. In general, it can be concluded that lead time for explaining emergent situations to a human and transitioning control should happen within a few seconds, while for non-critical situations, such as post-trip analysis, the amount of time may be as long as needed.

3) *All-time or only necessary explanations?*: It is also important to consider that humans need to enjoy their trips with AVs and get information from a vehicle when it is necessary. This aspect also applies to the delivery of explanatory information to end users. When the passengers/human drivers are provided with tons of information during the trip, it can lead to mental overload for them [281]. Consequently, it is generally favored that driving decisions and traffic scenes may be described to humans on board when traffic conditions are critical, and people need to be alerted. It is also noteworthy to specify that AVs must be equipped with need-based HMIs to deliver explanations. There are some challenges with effective automotive HMI design. First, people may have different choices or preferences for HMI (i.e., display monitor, alert interface, etc.). Furthermore, users' various cognitive and functional abilities must be a crucial factor in the design of user interfaces [15]. For instance,

people with visual or hearing impairments may need a customized HMI. Hence, automotive manufacturers must consider the diversity of users, contemplate the timing perspective of HMI explanations in line with relevant actions, and reach a consensus on the best practices with effective HMI design for AVs [234].

Based on the mentioned process steps and crucial elements, we see that achieving the interpretability of self-driving models is challenging, necessitating integration of those steps and cooperation between users and AVs. Consequently, while we argue that transparent and highly autonomous driving is feasible, human factors must be a vital consideration in the design and development of such systems. A simple graphical illustration of our proposed framework with its elements can be seen in Figure 3.10.

3.5 Research Gaps

As described in Chapter 3, AV explanations have been primarily investigated from algorithmic perspectives and presented in various forms and contents. However, these studies have a few general limitations. First, the reviewed literature does not put enough emphasis on human factors. As AV explanations are intended for target recipients, explanations must meet the users' inherent expectations and prior opinions from a sociotechnical perspective. In addition, these explanations are mainly non-interactive and do not enable users to engage in follow-up queries in presented information and judge real-time driving decisions. Finally, although the automotive community has broadly investigated various forms and contents of explanations, time granularity of explanations has not received considerable attention from researchers. Hence, there is a need to investigate the mentioned concepts to improve the explainability of autonomous driving systems. In this sense, in the following chapters, I design a series of experiments and perform analytical studies to bridge the gaps between missing pieces. These experiments present interactive explanations between humans and an AV, investigate the timing sensitivity of explanations for modular

and end-do-end autonomous driving, and describe critical challenges in promoting a human-centered approach to explainable autonomous driving research.

Chapter 4

Explaining Autonomous Driving Actions with Visual Question Answering

This chapter presents my preliminary experiment on explaining autonomous driving actions with a question-answering approach. Its content is primarily based on [17].

4.1 Introduction

When humans drive or are passengers on board, they inherently analyze real-time and upcoming traffic scenes and think about relevant causal, temporal, and descriptive questions, such as "Why is the car turning left?", "What action will the car in the left lane perform at the T-junction?" and "What is the speed of the vehicle in front?". Getting answers to such questions by any means helps us have a reliable and safe trip. In the context of autonomous driving, this problem can be formulated as a visual question answering (VQA) task so that an automotive user interface presents answers to user questions on action decisions and traffic situations. In this sense, I leverage the VQA mechanism to justify autonomous driving actions reflecting the car's decision-making in specific scenarios.

The framework is built as follows. I train an RL agent (i.e., an ego car) to operate in an autonomous driving environment and record its decisions (*actions*) in correspondence to the video frames (*states*). I then use a VQA system to justify actions

of the autonomous car: the VQA framework inputs an *image frame* with a *question* reflecting the action of the car in the scene and tries to predict the relevant answer for such an action.

Overall, the main contributions of this experiment can be summarized as follows:

- I present a preliminary empirical study on explaining autonomous driving actions with a VQA approach.
- I show that connecting vision and natural language could rationalize an RL agent's decision-making in an intelligible way.

The following sections describe the details of the experimental design, empirical results, and analysis of these results.

4.2 Experimental Design and Methodology

The proposed framework is designed in three primary steps. First, I use a deep RL agent to control an autonomous car in a simulation environment and collect a driving video from its field of view (FoV). I then convert this recorded video to image sequences at a uniform rate. Finally, I select five specific action categories in the extracted driving frames and annotate them using question-answer pairs that justify the car's action in the scene (Table 4.1). The high-level description of the components and overall architecture is provided in Figure 4.4. Given such a setup, the objective of our architecture is to predict the correct answer to a posed question about an autonomous car's performed action in an unseen driving scene. The details of the data collection, data annotation, and question-answering steps are described in the following subsections.

4.3 Data Collection

To obtain driving data, I train an RL agent (i.e., a self-driving car) on the CARLA simulator [76]. I use the DDPG algorithm [163] for the control of a self-driving car

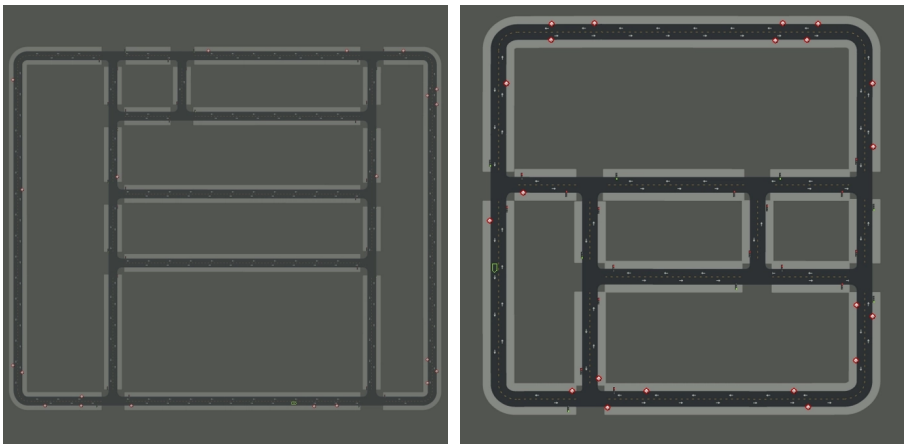


Figure 4.1: An aerial view of Town 1 and 2 on the CARLA simulator [76].

in a simulation environment. Control commands of automated driving have continuous actions including braking, acceleration, and steering angle which themselves can have a broad range of values as a representation. DDPG, as an augmented version of the Deep Q-learning algorithm, is particularly well-adapted for continuous action spaces and therefore is appropriate for driving control tasks. Furthermore, DDPG uses *experience replay*, a memory storing the agent's past experiences as *state, action, reward, next state* quadruples (s_t, a_t, r_t, s_{t+1}), out of which the algorithm can sample randomly to train the agent. This ability to reuse samples makes DDPG a computationally efficient learning approach. Moreover, DDPG has an actor-critic architecture, in which the actor learns an observation-to-action mapping, and the critic learns to evaluate the quality of an agent's chosen actions. DDPG also uses *target networks* - the target actor network μ', and target critic networks Q'. These networks are time-delayed copies of their original networks that help stabilize the training process. The parameters of target networks are updated as follows:

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{4.1}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{4.2}$$

where $\tau \ll 1$. For an effective action exploration, the term additive noise is usually added to the exploration policy and action is selected accordingly:

$$a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t \tag{4.3}$$

Such a learning technique enables the DDPG agent to learn a policy that maximizes its expected reward while also considering the quality of the chosen actions.

RL Training Details: I generate driving data by training the agent on Town 1 within CARLA. Town 1 (see Figure 4.1, a) is a map containing straight lines, left turns, right turns, T-junctions, traffic lights, speed signs, and various stationary objects around the curbs. I first use the A* motion planning algorithm [105] to generate a route with an initial and final point of a motion trajectory inside the simulated town, which shows consecutive waypoints linking these points. In our experiment, I set the number of waypoints to 15. By default, the waypoints are referenced to the origin point (0,0,0) in the map. To ensure that they are referenced to the dynamic position of the self-driving car while in motion, I use Pérez-Gil et al.'s methodology [211] and apply a transformation matrix to represent the state of the agent with these points, the vehicle's yaw angle, and its global position on the map as follows:

$$\begin{bmatrix} \cos\phi_c & -\sin\phi_c & 0 & X_c \\ \sin\phi_c & \cos\phi_c & 0 & Y_c \\ 0 & 0 & 1 & Z_c \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.4}$$

The goal of the task is that the ego car follows this predefined route and reaches the final destination by performing the relevant actions along its trip.

As seen from Figure 4.2, the agent acquires a driving vector f_t= (v_t, d_t, ϕ_t) from the simulation environment where these parameters reflect the vehicle's velocity, lateral distance, and yaw angle, respectively. Ideally, the goal of driving is to move on in the direction of the lane as long as possible without lane departure and collisions. In this

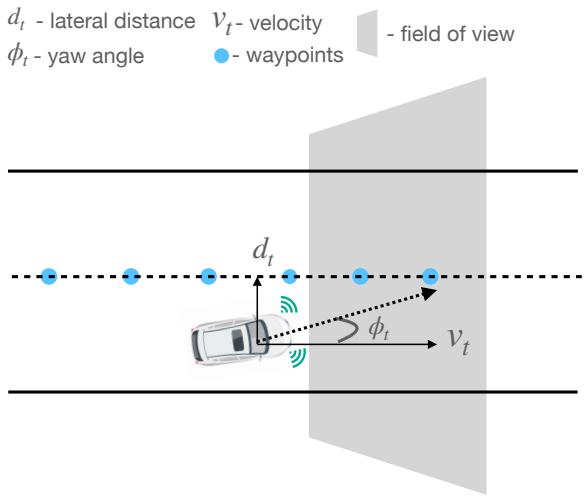


Figure 4.2: State space representation of the ego car in the driving environment. The ideal driving state is that the vehicle follows the direction of the lane within the lane.

sense, the reward shaping can be conditioned for the vehicle's 1) perfect longitudinal direction, 2) deviation from the lane direction with yaw angle, and 3) lane departure and collision. Based on these criteria, we adopt the relevant reward formulation from Pérez-Gil et al. [211] for an ego car:

$$R = \begin{cases} -200 & \text{road departures or collisions,} \\ \sum_t |v_t \cos\phi_t| - |v_t \sin\phi_t| - |v_t|\,|d_t| & \text{driving inside the lane,} \\ 100 & \text{arriving at the goal position.} \end{cases} \tag{4.5}$$

Finally, action space is continuous and can receive values from the interval [-1,1]. By defining this setting, I train the agent in Town 1. The training parameters of DDPG can be seen in Table 4.2.

4.3.1 Data Annotation

Once obtaining the driving video with the DDPG agent, I select 5 action categories (*go straight, turn left, turn right, turn left at T-junction, and turn right at T-junction*), and extract consecutive frames uniformly (30 frames per second) for 5 video segments. I then choose 10 frames from each segment. I especially ensure that these frames are extracted from driving segments, where the car follows the predefined

Table 4.1: Annotated question-answer pairs in our VQA framework

Action category	**Question**	**Answer**
Go straight	Why is the car going straight?	Because the road is clear.
Turn left	Why is the car turning to the left?	Because the road is bending to the left.
Turn left at T-junction	Why is the car turning left at T-junction?	Because there is no obstacle on the right side and turning left can be performed safely.
Turn right	Why is the car turning to the right?	Because the road is bending to the right.
Turn right at T-junction	Why is the car turning right at T-junction?	Because there is no obstacle on the left side and turning right can be performed safely.

route and perform the relevant action safely without lane departure or collision. I distinguish left and right turns in the current line from left and right turns at T-junction, as in the latter an ego car also has an alternative route. So, the training data includes 5 action categories with 50 high-quality frames per category, denoting a total of 250 driving scenes obtained from the recorded video. I manually annotate each instance of the training data with a relevant question-answer (QA) pair (see Table 4.1). As test data, I select a collection of 100 frames from both Town 1 and Town 2 on the CARLA simulator, as the map of Town 2 is similar to Town 1. Similar to the training data annotation, I select 20 frames for each action category and annotate them with relevant QA pairs. The goal is to assess the generalization ability of the employed VQA framework on these action categories in unseen traffic scenarios.

4.3.2 Question-Answering Framework

On the question-answering side, I fine-tune the original VQA framework [12] trained on the MS COCO dataset [166]. At the highest level, the VQA model takes an

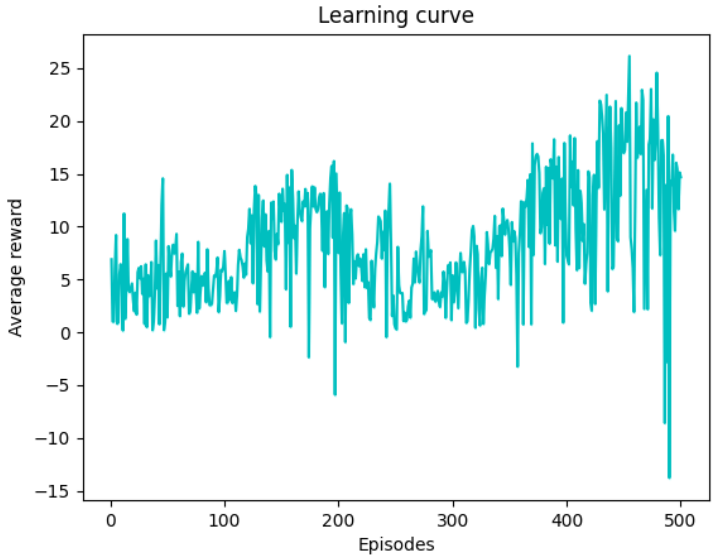


Figure 4.3: Learning curve of DDPG in Town 1 with the specified parameters. The proposed VQA framework is further fine-tuned on driving data collected here.

encoded driving image and a question embedding as input, to predict the answer (i.e., explanation) for a performed action in the scene. The model is composed of two neural networks. The first one is a multilayer-feedforward network with 2 hidden layers each containing 1000 hidden units and uses a *tanh* activation function. I apply the dropout regularization with 0.5 in each layer. Finally, a long short-term memory (LSTM) [114] followed by a softmax layer is employed to produce an answer for the asked question about the driving action. On the image encoding part, I eliminate the output layer and use the last hidden layer of the pre-trained VGG-19 architecture [246], producing a 4096-dimensional feature vector. Further, a linear transformation is applied to make the image features 1024-dimensional. The LSTM model for the question encoder has 2 hidden layers with 512 hidden units, and thus it is a 1024-dimensional vector, the same as image features. An interesting aspect is the unification of the question and image vectors from a mathematical perspective. Previous studies have generally either preferred the concatenation or element-wise multiplication of these vectors, but [12] have empirically shown that multiplying the image and question encoder usually

Table 4.2: The training parameters of DDPG on CARLA

Actor learning rate	Critic learning rate	Target network hyper-parameter	Replay buffer size	Batch size	Discount factor
0.0001	0.001	0.001	100000	32	0.99

leads to a better representation. Consequently, given the image vector, V_i, and question embedding V_q, the resulting vector passed to the fully connected layer of the VQA pipeline is represented as their element-wise multiplication, as a fused feature vector:

$$V_r = V_i \times V_q \tag{4.6}$$

I use the question and answer vocabularies of the original VQA framework, which have sizes of more than 17K unique tokens and 1000 candidate answers (which are either single tokens such as "yes," "white," or expressions consisting of two or more strings such as "playing video game"), respectively, obtained by descriptions from the MS COCO [166] images. I customize candidate answers by adding our answers of 5 action questions to that answer vocabulary. The expectation is that the VQA model picks the most correct answer with the highest softmax probability score out of the 1K candidates for the asked "Why" question about the action within the driving scene. Figure 4.4 describes the proposed framework: Driving video from the vehicle's field of view, collected from the CARLA simulator, is converted to image sequences at a uniform rate. Furthermore, the fine-tuned VQA model inputs a driving scene image and textual question reflecting an action in that scene and predicts the answer for the asked question on this action out of the 1K possible answer set.

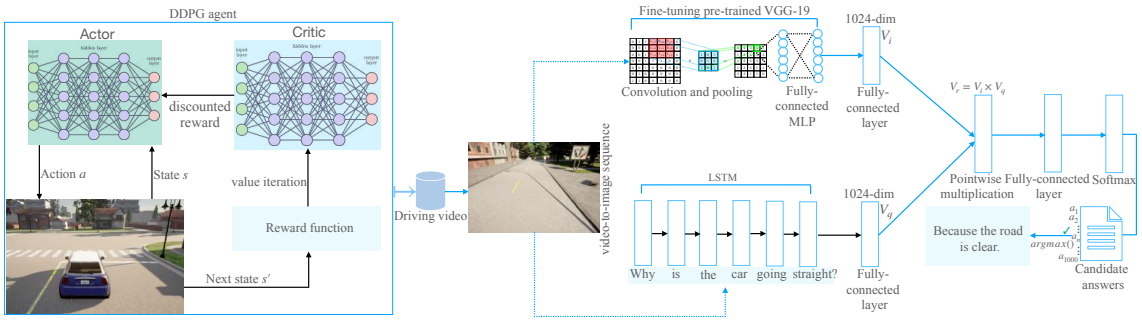


Figure 4.4: A diagram of the proposed VQA architecture for autonomous driving.

4.3.3 Experimental Results

On the data collection side, I train the DDPG agent on the CARLA 0.9.11 version in 500 episodes using a TensorFlow backend to get a driving video. As described above, I use 250 frames from Town 1 for training our VQA network and evaluate its performance on 100 frames collected from Town 1 and Town 2 (Figure 4.1). I use the PyTorch backend for training and evaluating our VQA architecture. The experiments have been performed on an NVIDIA RTX 3090 GPU machine with a 32 GB memory size. All the frames have been set to have a size of 640 $\times$ 480 both in training and test. As there are ground-truth answers (i.e., the "Answer" column in Table 4.1) for the asked question about an image, I compare the top prediction of our model on the test data (i.e., an answer with the highest softmax probability score) with these ground-truth answers. Thus, I use accuracy as an evaluation metric, which is defined as follows:

$$\text{Accuracy} = \frac{\#\ frames\ with\ correct\ predictions}{total\ number\ of\ test\ frames} \tag{4.7}$$

Based on this evaluation criterion, the proposed VQA model predicted 80 correct answers to the asked questions for 100 images. Hence, the accuracy of the prediction is 0.8 or 80%.

Discussion: Except for *turn left* actions, the model predicts explanatory answers correctly and confidently (i.e., see average softmax probabilities in Figure 4.6) for

Table 4.3: Number of correct predictions for each action category

Go straight	**Turn left**	**Turn left at T-junction**	**Turn right**	**Turn right at T-junction**	**Total**
20/20	0/20	20/20	20/20	20/20	80/100

all remaining action classes. Interestingly, in the frames with turn-left scenarios, the VQA framework primarily recognize these actions as *turn right*. In Figure 4.5, we provide exemplary driving scenes for the five action categories. As seen, the model was able to predict the highest probability scores for all actions in the scenes correctly, except for the misclassified *turn left* action in the second image. This misclassification could be due to ambiguity in the tested driving frames, the shape of curves in the scene, and road conditions in the training data. Hence, it is important to increase the size of the training data considering the shapes of road lanes and curves, lighting, and other road objects to potentially improve the accuracy of the predictions of the VQA network on driving actions.

Another implication of our work is that unifying computer vision with a natural language provides an opportunity to explain temporal actions of an RL agent. As explored in a recent study [210], explaining RL in sequential decision-making problems is an important and emerging topic, particularly when explanation receivers do not have a technical background. As autonomous driving is a safety-critical application area, justifying reinforcement learning-based decisions to end users with natural language-based reasoning is an effective and easily understandable approach. In general, an RL agent's interaction with the environment as an MDP can be implemented as model-free or model-based RL. A natural foundation for explainable reinforcement learning (XRL) would be to provide reward or policy-based justifications for action decisions. To this end, model-based RL is a promising approach to XRL as the agent first tries to understand the world with *prior knowledge* and then develops a *model* to represent this world, where the approach is called *planning*. The planning process uses a model representation to generate a predicted future trajectory [255, 289].

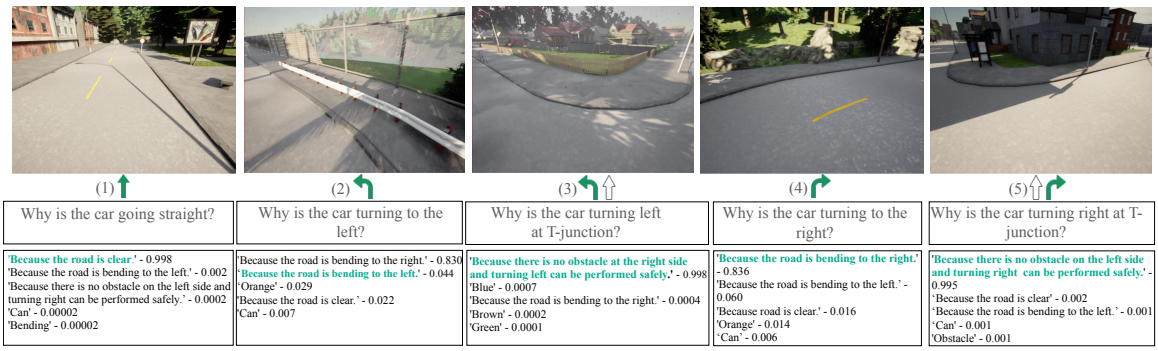


Figure 4.5: Example scenarios from an ego vehicle's field of view on CARLA. During the decision-making process of the agent, we are given visual signals and we ask action-related questions and try to find an answer given the current state. The green arrow shows the ego car's chosen action and the white arrows indicate the other route at T-junction scenarios. We show the top 5 answers predicted by our model. The green-colored text shows the correct answer to the question for the performed action of the car. Except for the *turn left* scenario, justifications for other actions are predicted correctly by the model.

According to the model projection, an optimal action is decided at each planning step, which provides a predicted state and a predicted reward. The predicted states and rewards can be analyzed and visualized for the planned trajectory, thus providing an explanation of why the agent favors the choice of a particular action at a specific time step from an algorithmic perspective. However, as self-driving explanations are mainly targeted at the general community, it is essential to ensure that explanations delivered to the users are intelligible and informative for them. While [210] has attempted to build an inherently explainable RL architecture, I build the explanations independent of an agent's decisions. I also acknowledge the need to be cautious about providing explanations that are independent of an agent's behavior; it is possible that post-hoc explanations may *not* always reflect an agent's actual decision-making process. For example, in an actual *left turn* scenario, a model's response to the question "Why is the car turning to the *right*?" as "Because the road is bending to the *right*." would be a hallucination of a VQA architecture. Consequently, it is important to further investigate the topic of generating linguistic explanations for an RL agent's actions and evaluate such explanations with human adversarial examples as well.

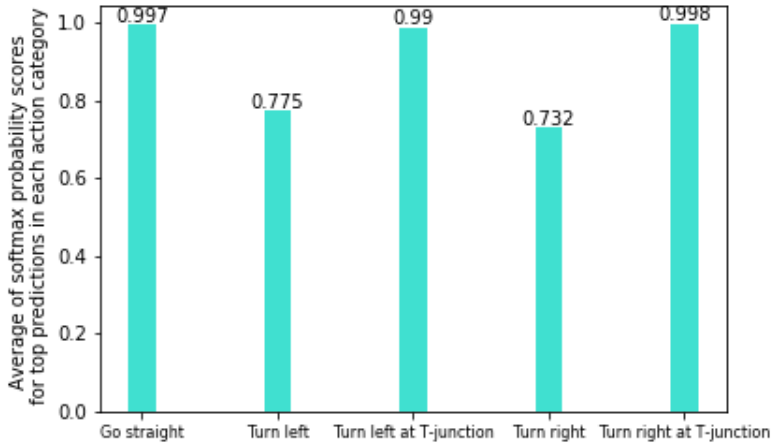


Figure 4.6: The average softmax probability scores for top predictions in each action category.

Limitations of the experiment: Real roads are more complex and dynamic with the presence of traffic lights, bystanders, passengers, other vehicles, and adverse weather conditions. In the current version of our framework, the ego car only interacts with the stationary environment and explains actions associated with such interactions. Moreover, the dataset is small in size. Consequently, I use these limitations as a motivation for further direction and overcome these issues with the experiments presented in subsequent chapters.

Practical use cases: In practice, the VQA mechanism can be leveraged at least in two ways on real autonomous cars. First, it can help passengers on board monitor driving safety by "judging" the vehicle's decisions. For instance, a user interface or dashboard set up on a back seat may provide voice-to-text functionality, and a passenger can observe driving surroundings, ask a question about the vehicle's chosen action, and get an answer. Such a feature can help monitor the reliability of self-driving and instill trust in vehicle autonomy during the trip. Another practical application is to retain a history of action-question-answer triplets $(...a_t, q_t, ans_t, a_{t+1}, q_{t+1}, ans_{t+1}...)$ and use it for forensic analysis in possible accident investigations with self-driving

vehicles: Recorded explanatory log data can help understand why the self-driving vehicle made a specific decision at a particular time just before being involved in an accident.

4.4 Summary

This chapter has presented a preliminary study on explaining autonomous driving actions with a VQA approach. I use driving data generated by an RL agent on the CARLA simulator and develop the question-answering system as an explanatory approach to the agent's decisions. The experimental results show that a simple and straightforward VQA mechanism can help interpret the real-time decisions of an autonomous car and also help understand its correct and incorrect decisions as safety implications. The results also suggest that unifying VQA with RL-based decision-making will likely do well for actions in a dynamic environment, provided that we have rich training datasets. In this sense, in the subsequent chapters, I perform experiments in more dynamic environments and describe opportunities and challenges with knowledge-based QA models for capturing traffic dynamics.

Chapter 5

Transition to Large Pretrained Model-based Explanations: A Paradigm Shift

5.1 Introduction

Pretrained large language models (LLMs) and vision-language models (VLMs) have recently revolutionized NLP, robotics, and computer vision across many tasks. Primarily based on BERT [70] and GPT [218] families, there has been a surge in building general-purpose language models. These models are called *Foundation Models* where their internal weights are further adjusted (i.e., "fine-tuned") with a domain-specific knowledge base for serving task-specific purposes. In this context, Foundation Models have been an area of interest for interpretable autonomous driving since the year 2023. The advantage of these models is their massive internal knowledge, which has the potential to capture the semantics in dynamic environments and sequential decision-making problems. This feature is strongly relevant to interpreting autonomous driving actions. The visual comprehension ability of Foundation Models can help provide human-interpretable descriptions of traffic scenes and the behavior of a self-driving vehicle. These pieces of information can be textual descriptions, visual explanations, or other forms of reasoning techniques, depending on the task.

5.2 Large Language Model and Vision-Language Model-based Explanations for Autonomous Vehicles

The role of LLMs and VLMs in explainable autonomous driving has recently been investigated from several aspects [57, 303]. For example, the Talk2BEV [50] model has incorporated a VLM into bird's-eye view maps, enabling spatial and visual analysis to predict unsafe traffic scenarios. In a similar work, DriveGPT4 [286] justifies an autonomous car's decisions via textual descriptions and responds to humans' questions in the same manner. The concept of question-answering as a human-machine interaction has been explored in two further models LingoQA [179] and VLAAD [205]. Other recent prominent large pretrained models for explainable autonomous driving include LINGO-1 for producing live natural language explanations [277], Driving with LLMs [44] and Drive Like a Human for context understanding [94], GPT-Driver for interpretable motion planning [178], and DILU [279] and LanguageMPC [241] for human-like decision-making.

Overall, generalization and emergent abilities of large pretrained models is a new paradigm shift for combining vision, language, and action and providing interpretability across many tasks in control problems. This success has resulted in the emergence of a new topic termed *Embodied AI* where an agent learns and improves its performance continually by interacting with the environment with its massive internal knowledge [78, 275]. Perhaps there will be more work to improve the explainability of Embodied AI in the context of autonomous driving as such models can capture the semantics and temporal changes in traffic scenes. However, the current well-known limitations of these large models (i.e., "hallucinations," authoritative but factually incorrect responses) can have dire consequences for driving safety in real, physical environments, and damage user trust in AVs. Chapter 6 sheds more light on these caveats and describes the implications of VLM-generated hallucinations on users'

Table 5.1: Studies on large language models and vision-language models-based explanations for AVs

Study	Task	Algorithms/Methods	Delivery format	Target audience
Dewangan et al., [50], 2023	Language-augmented Bird's-eye View Maps for Autonomous Driving	GPT-4, GRIT	Visual, Textual	All groups
Xu et al., [286], 2023	VQA and natural language-based explanations for autonomous driving	LLAMA 2, CLIP	Visual, Textual	All groups
Marcu et al., [179], 2023	Video question answering for autonomous driving	Vicuna-1.5-7B, GPT-4	Textual	All groups
Chen et al., [44], 2023	Improving context understanding in autonomous driving with object-level vector modalities and LLM	GPT 3.5, PPO	Visual, Textual	All groups
Fu et al., [94], 2023	Understanding traffic situations in a closed loop	GPT 3.5	Textual	All groups
Mao et al., [178], 2023	Interpretable motion planning as language modeling	GPT 3.5	Textual	Road users
Sha et al., [241], 2023	LLM as a decision-maker in complex driving scenarios	ChatGPT, MPC	Visual, Textual	AV developers
Wayve Team [277], 2023	Providing live explanations in natural language	Integrated vision, language and action architecture	Textual	All groups
Nie et al., [197], 2023	Interpretable reasoning in complex driving situations in autonomous driving	GPT-4, MLP, ViT-G/14	Textual	All groups
Park et al., [205], 2024	Video question answering for traffic scene understanding	Video-LLAMA, GPT-4	Textual	All groups
Wen et al., [279], 2024	LLM-based knowledge-driven approach for interpretable autonomous driving	Out-of-box LLM	Visual, Textual	AV developers
Yuan et al., [292], 2024	Retrieval-augmented VLM for explainable autonomous driving	LanguageBind, MLP, Vicuna-1.5-7B	Textual	All groups
Chi et al., [46], 2024	GPT-aided explainable decisions for autonomous vehicles	GPT, Graph of Thoughts	Textual	All groups
Atakishiyev et al., [19], 2024	Robustness of a transformer-based VideoQA model against human-adversarial questions and its safety implications for self-driving	Video-LLaVA	Textual	All groups
Duan et al., [79], 2024	Unifying imitation learning with LLMs to enhance safety of end-to-end driving	Vicuna LLM, Swin transformer	Textual	All groups

trust, perceived safety, and feeling of comfort with AVs. A summary of studies on LLM and VLM-based explanation approaches for AVs is described in Table 5.1.

Chapter 6

Incorporating Explanations into Human-Machine Interfaces for Trust and Situation Awareness in Autonomous Vehicles

This chapter describes my situation awareness framework for autonomous driving. It further presents an empirical study on this framework and validation of the results with human judgment. The chapter is primarily based on [18].

6.1 Introduction

While XAI is becoming essential in designing and developing modern AVs, another crucial factor is the effective delivery of these explanations to the users in the loop. Consequently, automotive HMIs of various types must be built in to display driving information appropriately to users. I provide an example of such interaction in Figure 6.1: Waymo's self-driving vehicle provides explanatory information to the pedestrians in front and the human drivers behind via a simple interface supplied on top of the vehicle. While the pedestrians can interpret this information as "Waymo is yielding to us and is not going to drive while we cross," the human drivers at the rear may perceive the intent signal in the back view as "Waymo is stopping because it is yielding to the pedestrians crossing the road." Such a simple explanation provision

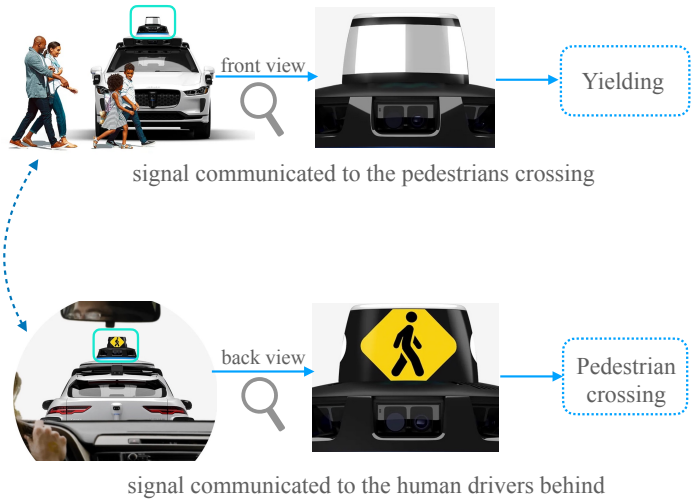


Figure 6.1: An example of explanation communication to the pedestrians (top) and a human driver at the rear (down) by Waymo's self-driving car via its external HMI. The green bounding boxes have been manually added to indicate these signals. The figure has been drawn based on the content in [11] and [276].

method shows that HMI can effectively describe traffic scenes to road users and help them be aware of the self-driving car's behavioral intention.

Considering the immense need for AI transparency and the role of HMI for possible control and self-driving monitoring, this chapter contributes a structure to the combined role of XAI and HMI for trustworthy autonomous driving. By definition, as XAI aims to provide interpretable models while maintaining good model performance and enabling users to comprehend, trust, and manage the intelligent system, we investigate all these aspects in our study. I specifically analyze the problem setting with a "3W1H" approach: *what* information to deliver, *when* to deliver, *whom* to deliver, and *how* to deliver explanations using a supplied user interface. In this sense, I systematically review prior investigations and reveal the suggestions and the practical recommendations derived from that literature. As a result, I incorporate insights from these studies into a framework to improve situation awareness and trust in AVs.

Overall, the contributions of this chapter are threefold:

- An investigation of prior studies that explore XAI and HMI for autonomous driving and identification of the best practices within the "3W1H" approach;
- An HMI and XAI-guided situation awareness framework for autonomous driving and experiment on the framework;
- A user study and validation of the experimental findings with human judgment.

6.2 The "3W1H" Aspects of Explanation Conveyance to End Users

The design and use of human interfaces have been of interest to the automotive community since the development of autonomous driving. Human-centric interactive system design does not solely provide driving-related information to users but also can help with meeting their individual needs during a trip with a self-driving vehicle [234, 238]. In this regard, there are several challenges in the communication of explanations about driving decisions to end users. First, explanations should be relevant to a user's mental model: as user satisfaction is the target focus in the development of autonomous driving systems, the content of explanations should meet their needs [281]. Furthermore, which users (drivers, passengers, external observers) could benefit from these explanations is another aspect of targeting explanations. In addition, the timing mechanism of explanations is also an essential property of the user interface to ensure that the self-driving vehicle's actions or driving scenes are described within appropriate time frames [281]. Finally, how explanations are delivered is also a crucial factor in meeting users' expectations. Considering these four-dimensional properties of explanations in conjunction with HMI, I explore this topic systematically and present the "3W1H" approach.

6.2.1 What?

What types of explanatory information are needed for road users? Explanations can describe the stationary and dynamic objects in the scene and inform users of how an autonomous vehicle perceives these objects during a journey [144]. An interactive interface inside a self-driving vehicle can provide updates on traffic situations, weather conditions, and offer customizable preferences (i.e., the temperature inside the vehicle and individual options depending on the users' needs). Moreover, the autonomous car can share information outside to bystanders and other vehicles as a part of the vehicle-to-vehicle (V2V) or vehicle-to-user (V2U) communications.

In addition, and as an ongoing research direction, practical XAI methods have recently been explored to describe the rationale behind the decision-making process of AVs in a human-interpretable format [20]. The critical aspect of explanation techniques is how satisfactory these approaches are from the users' perspective. In this context, users' mental models must be considered: users may favor specific types of explanations depending on the driving scenario. In general, it has been shown that explanations are more related to *traffic-related event cognition* and *describing the driving behavior of AVs* as an answer to the "What" question in this context [238].

6.2.2 When?

While a history of continuous action- and scene-explanation pairs can be recorded for possible post-trip driving analysis and forensic investigations, it is noteworthy to mention that users, particularly passengers and human drivers do not always need explanations. According to Koo et al.'s study [153], seamlessly delivered explanations can result in *mental overload* for human drivers in semi-autonomous driving and distract them from meeting their primary obligations. Instead, they can be alerted when an autonomous mode decision is initiated, when driving conditions change substantially, when takeover requests are made, and when sensor failure occurs. In another user study, Haspiel et al. [106] have focused on the significance of timing for ex-

planations, and their exploration concludes that human drivers favor explanations just *before* action is decided rather than receiving explanations *after* the action is performed. Avoiding a potentially overwhelming information flow also applies to explanation provision for passengers. Kim et al.s' [144] recent user study on a real road with a wizard experimenter reveals that users favor explanations in critical and risky traffic circumstances rather than having continuously presented behavioral information. Similarly, Shen et al. [243] have also qualitatively validated the premise that people favor explanations in near-crash and emergent situations.

6.2.3 Whom?

User categories encompass individuals who directly interact with a self-driving vehicle (i.e., passengers and backup drivers) or are affected by its presence and operation on roads [20]. Explanatory behavior can be conveyed to road users outside the vehicle (i.e., bystanders, pedestrians, and cyclists) using language or intent signals as well as shown in Figure 6.1 to ensure that individuals nearby remain aware of the behavior of the autonomous vehicle while it operates. Moreover, people outside and adjacent to an autonomous vehicle, such as bystanders, pedestrians at crosswalks, drivers operating non-autonomous cars, and cyclists, may also expect behavioral information from an autonomous car. Additionally, understanding decisions of a self-driving vehicle may be a necessity for traffic enforcement officials [102] and emergency responders [169]. To ensure operational safety, traffic enforcement personnel may carry out a compliance check on the self-driving vehicle at some time and the vehicle may be requested to provide some form of explanation on its motion. Furthermore, behavioral information may be helpful for emergency responders for effective response to the accident and emergency conditions caused by/with the presence of an autonomous car. Overall, explanations should be conveyed to relevant interaction partners in critical times/ whenever needed while an autonomous car operates on roads.

6.2.4 How?

Various studies have shown that users might have different preferences for receiving explanations depending on their identity and traffic situation. For example, Faltaous et al.s' [86] user study in a simulation environment shows that providing multimodal explanations with *auditory, visual*, and *vibrotactile* feedback is effective for potential takeover requests in highly-urgent driving conditions. Furthermore, Schneider et al. [236] have tested five various feedback techniques - *light, audio, object visualization, textual information*, and *vibration* - in a virtual driving scene, and evaluate user satisfaction with these explanation modalities. Their findings show that light or object visualization is preferred more for proactive situations, while sound and light are more favored for reactive scenarios. Interestingly, the users are not satisfied with textual descriptions or sound-based feedback and even find the latter disturbing in the long term. Detjen et al. [69] show that a *planar heads-up display* (pHUD) and *contact analog heads-up display* (cHUD), as an augmented reality presenter, have been liked by testers as an effective display technique in both low and highly automated driving settings. In contrast, Dandekar et al. [62] have investigated an effective way of conveying driving information to passengers while not distracting them from their in-vehicle activities. The users' feedback indicates that a colored light bar display and windshield display are satisfactory interfaces and do not disturb them while being immersed in non-driving related infotainment. Kim et al. [144] also conclude that the semantic segmentation map of objects displayed via windshield display instills trust in human participants in the behavior of an AV.

It is also noteworthy to state that automotive HMI design should consider humans' varying physical and cognitive capabilities and be inclusive of everyone. Particularly, people with some form of physical and cognitive impairment may need customized user interfaces. In recent research, Arfini et al. [15] have presented theoretical and pragmatic challenges in the effective interface design for people with some functional

Table 6.1: The "3W1H" aspects of explanation conveyance to autonomous driving users based on the findings of prior studies

Dimension of explanation	Description
What?	Decisions of an autonomous vehicle, traffic scenes, and events
Whom?	Passengers, human drivers, people with cognitive and physical impairments, remote operators, bystanders, cyclists, traffic enforcement officials, emergency responders
When?	Critical and emergent situations, takeover scenarios, the time before an action is performed
How?	Audio, visual, vibrotactile, text, heads-up display, passenger intervening interface, haptic feedback, braille interface

limitations. Autonomous vehicle users may have visual, hearing, mobility, and speech impairments, and customized user interfaces may help them get explanatory information on the vehicle's actions and traffic situations. For instance, a braille interface can help people with visual difficulties to acquire driving-related information from the vehicle. Or a gesture recognition system may be implemented to track the user's hand movements and enable them to interact with the vehicular interface. Hence, in general, we can summarize that no single type of HMI can meet the needs of all autonomous car users due to their diverse cognitive and physical abilities and subjective preferences.

While the "3W1H" approach reveals a broad spectrum of explanation conveyance to end users (see Table 6.1), the goal remains the same: bringing situation awareness to people inside and outside of an autonomous car. In the next subsection, I present a unified approach to a general situation awareness framework via XAI and HMI and describe the implications of such a framework for people in the loop.

6.3 A Unified Approach to Situation Awareness Framework via Explainable AI and Human-Machine Interfaces

The goal of an XAI system is to provide explanatory information on the system's particular decisions or behavior. According to the definition of the term, *situation awareness* by Endsley [83], users in the loop must be aware of what particular decision the AI system made, why it made that specific decision, and what decisions it will make in the next similar state at a later time. The three concepts have been referred to as *perception*, *comprehension*, and *projection*, and are viewed as indicators of situation awareness in human-in-the-loop AI systems. Sanneman and Shah [233] have further expanded that framework from an *informativeness* perspective. They correspond the aforementioned concepts into three levels of explanations - Level 1, Level 2, and Level 3 - for human-centered XAI systems, respectively. Here we adapt Sanneman and Shah's [233] framework to autonomous driving scenarios and describe the implications of such a framework for self-driving vehicle users.

Given a traffic scenario $s \in S$, we denote a set of possible actions (i.e., go straight, turn left, etc.) that may be taken by an autonomous car as

$$A = \{a_1, a_2, ..., a_n\}.$$

At each time step of motion of the car, an explanation interface may describe what action the car is performing and why it is doing so. We can describe this particular action as $a \in A$ and the causal factor (i.e., traffic light, pedestrians crossing, etc.) that made the car behave in that particular way as $f \in F$. The explanatory information I_e for situation awareness can then be described via a combination of *four elements*: the traffic scenario, chosen action, causal factor inducing that action, and explanation delivery time:

$$I_e = (s, a, f, t) \tag{6.1}$$

This explanation is communicated to autonomous car users using various built-in

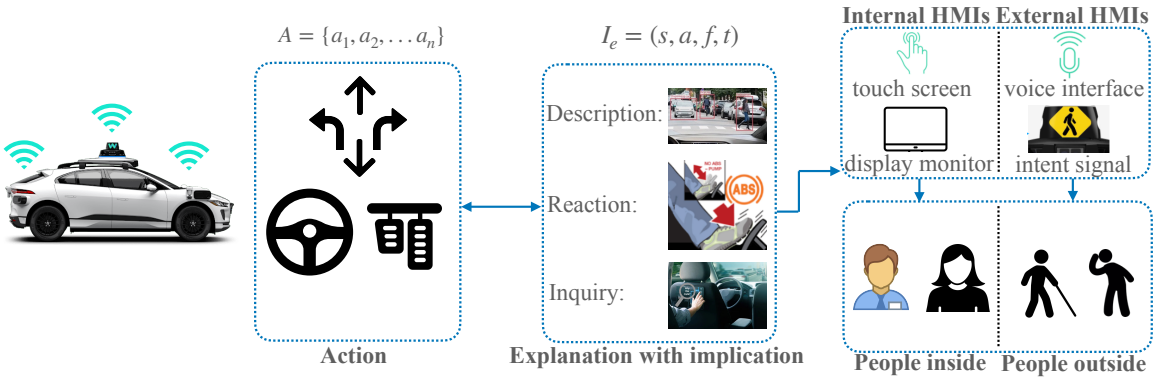


Figure 6.2: The proposed situation awareness framework for inside and outside users of an autonomous vehicle with XAI and representative HMIs

internal and external HMIs as needed. Depending on the interplay between the user and the explanation, the information conveyed can be classified as interactive or non-interactive, which can be stated as follows:

$$I_e = \begin{cases} \textit{interactive}, & \text{if reactive or inquisitive} \\ \textit{non-interactive}, & \text{if descriptive} \end{cases}$$

Now, let us define what we mean by *descriptive, reactive*, and *inquisitive explanations* in this context:

1. Descriptive explanations: A user interface provides general information about the action of the self-driving car and the scene it is moving through. The people inside or outside become aware of the driving environment and behavior of the self-driving vehicle without further interference.

2. Reactive explanations: The user interface invites the passenger or human driver to react to an emerging situation. For passengers, this could be an emergency override scenario where vehicular automation by design allows them to take control of vehicles in case other outreach actors can not control the vehicle. For human drivers, these are typically takeover requests, where the explanation interface communicates upcoming emergent states, and the driver must take control of the vehicle from the automated mode and manage the situation safely.

3. Inquisitive explanations: While descriptive and reactive explanations are al-

ready well-known to the automotive community, to the best of my knowledge, inquisitive explanations have not been investigated well in the current literature. Inquisitive explanations refer to the explanations that users, such as passengers, can ask the system to follow up on the previous response or to test the robustness of the interactive user interface. These questions can be anything related to the autonomous vehicle's decisions or traffic scene in general, but can also be "tricky" or *adversarial* questions aiming to stress test the explanation interface. For instance, assume in an actual *right turn* scenario under a green light, a passenger asks the conversational user interface, "Why is the car turning to the *left?*" as an adversarial question. In this case, the user interface should provide an action-reflecting response like "No, the car is turning to the right as the traffic light allows a right turn," explaining what the autonomous car is doing and why it is doing so. These questions may be asked deliberately or unintentionally (e.g., people with visual impairment may have difficulty perceiving the scene correctly) to test the trustworthiness of the vehicle, its actions, and its awareness of the surroundings. As a consequence, a human-machine interface should not only provide conventional explanations but also defend against potential adversarial queries to ensure that explanations are scene and action-reflecting. This feature is an essential property of automotive HMI. The ability to provide correct explanations can instill trust in users and encourage them to continue to use AVs. The graphical description of this framework is shown in Figure 6.2. With this approach, the proposed situation awareness framework may achieve Level 1 (chosen action) and Level 2 explanations (explanation for the chosen action), i.e., perception and comprehension, as described by Sanneman and Shah [233]. The authors describe Level 3 explanations as "XAI for projection" and relate it to counterfactual analysis, such as what an autonomous car will do if the driving scene changes in a certain way. Perhaps in the context of autonomous driving, Level 1 and Level 2 explanations are more essential for the users as they are more concerned about the actual scene the vehicle is getting through. With that said, curiosity to assess automation ability and

the reaction of the vehicle to further emergent driving conditions may necessitate Level 3 explanations for foreseeable situation awareness as well.

6.4 Case Study: Interactive Dialogues between a User and An Autonomous Vehicle

To validate our proposed framework, we perform an initial and simple empirical study and use human judgment to evaluate the effectiveness of the framework on the chosen traffic scenarios. As HMI can be both interactive and descriptive from an explanation communication perspective, we choose to design interactive explanations so that a user can prompt their question and expect context-aware information from an explanation interface. We formalize this setting as a *visual question answering* (VQA) problem. VQA is a learning task at the intersection of computer vision and NLP that inputs an image i, a textual question q associated with the content of the image, and predicts an answer a for the asked question. The problem can be more precisely represented as a tuple of the specified parameters:

$$x = (i, q, a)$$

Most of the state-of-the-art VQA models input an image and question parameters, then output a combined representation of vision and language $r \in \mathbb{R}^{d_r}$ via a multi-modal learning network f:

$$r = f(i, q)$$

Asking a question about the behavior of a vehicle is driven, at least partly, by human intuition. Becoming aware of traffic events and a human driver's decisions in the car helps passengers on board understand ongoing situations and have a comfortable trip. This intuition also can be related to asking meaningful questions to the conversational HMI of an autonomous vehicle about the driving scenes. In my preliminary study,

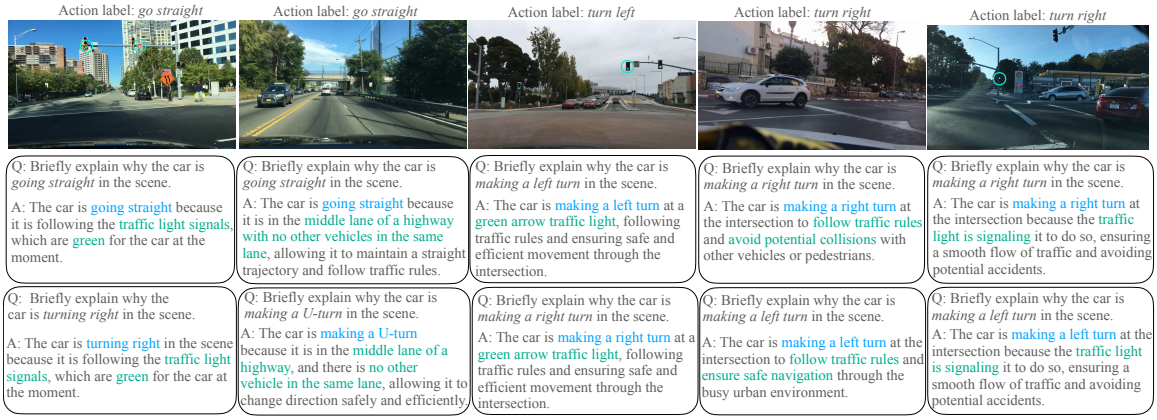


Figure 6.3: Our experiment on the five chosen traffic scenes from the BDD-A dataset with the LLaVA multimodal transformer. While LLaVA seems to yield correct explanations on *conventional questions* (top) with actual actions (blue-colored text) + causal factors (green-colored text), it fails to generate factual explanations on the *adversarial* questions (bottom). The bounding boxes have manually been added to indicate causal factors inducing the chosen actions.

I have shown the applicability of question-answering-based reasoning to explainable autonomous driving [17]. A crucial aspect of such interaction is the delivery of context-aware and action-reflecting explanations in a timely manner.

Experiment: Based on this motivation, I design the experiment as follows: I use Large Language and Vision Assistant, LLaVA [168], as a multimodal learning framework that inputs vision and textual query for describing driving scenes. LLaVA is a multimodal transformer architecture built on top of the Vicuna LLM [47] and the pre-trained CLIP visual encoder ViT-L/14 [219]. Given $\mathbf{X}_\mathrm{v}$, as an image, and $\mathbf{Z}_\mathrm{v} = g(\mathbf{X}_\mathrm{v})$, as a visual feature, LLaVA applies a trainable projection matrix $\mathbf{W}$ for converting the visual feature, $\mathbf{Z}_\mathrm{v}$, into language embedding tokens $\mathbf{H}_\mathrm{v}$ as follows:

$$\mathbf{H}_\mathrm{v} = \mathbf{W} \cdot \mathbf{Z}_\mathrm{v}, \quad \text{where} \quad \mathbf{Z}_\mathrm{v} = g(\mathbf{X}_\mathrm{v}) \tag{6.2}$$

I sample driving scenes from the real driving videos of the Berkeley DeepDrive Attention (BDD-A) dataset [284], and select five different scenarios shown in Figure 6.3. I then consider two types - *conventional* and *adversarial* - questions for the actions performed in these scenes. While the former is context-aware and action-reflecting,

the adversarial question refers to asking an *incorrect* questions about actions, deliberately or unintentionally, to stress test the robustness of the explanation provision model. The key idea with this test is that HMI must always communicate correct explanations and defend against human adversarial questions by providing context-aware information.

I present the explanations (i.e., responses) generated by LLaVA on the action-reflecting and adversarial questions in Figure 6.3. As seen, the model generates reasonable and human-interpretable explanations for conventional questions and justifies its responses. However, a tricky question easily confuses the model and makes it generate an incorrect answer to the asked question. For instance, in the leftmost scene in Figure 6.3, the car moves straight because of the green light (i.e., inside the left bounding box). Furthermore, there is a sign (i.e., inside the right bounding box) that *prohibits* the car from *making a right turn* at that intersection. Ideally, when we ask that adversarial question about the right turn, we might be satisfied with an explanation like "No, the car is going straight, and no right turn is allowed in the scene as the traffic signal prohibits it." However, we observe that the model is flawed with such a deliberate question and fails to present an adequate answer. The higher influence of a textual question on a VQA model's prediction is generally known as the *language prior problem*. Therefore, the possibility of adversarial questions must be considered in the construction mechanism of automotive VQA models, and relevant HMIs must provide corrective and context-aware information against such questions.

We need to consider that an autonomous car is equipped with an interactive HMI so that people inside the vehicle can ask such interactive questions at some point in the trip with the vehicle. To understand the impact of the generated responses on the users' feeling of comfort and safety perception in autonomous driving, we conduct a human study with the above experimental framework, considering them as passengers of a vehicle.

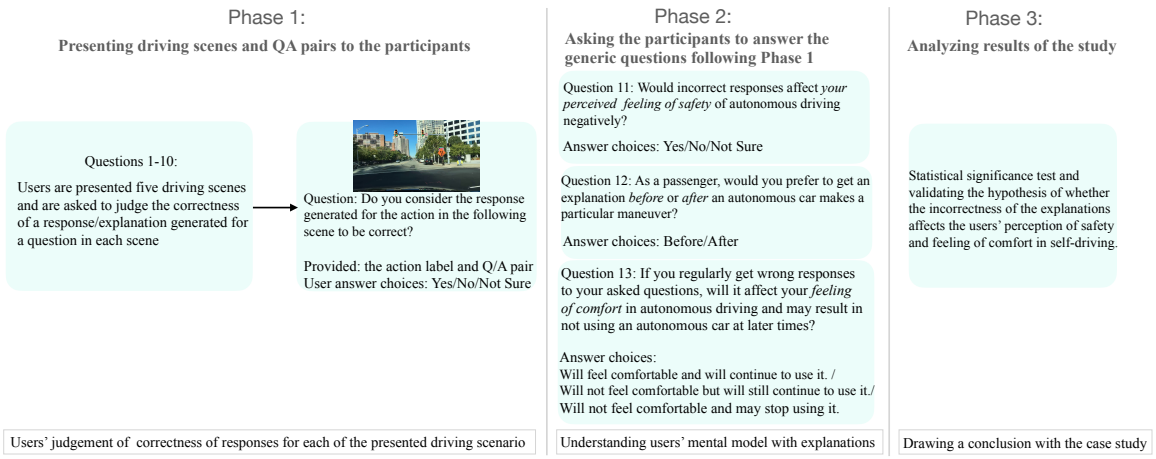


Figure 6.4: Design of the user study based on the experiment in Figure 6.3. The participants judge the correctness of explanations for each of the five scenes presented. After getting experience with explanations, they are asked two more questions on their perceived safety and mental comfort with the role of explanations while using an autonomous vehicle. Users' responses are validated with a statistical significance test to draw a conclusion with the case study.

6.4.1 Design of the User Study

I have performed a user study[1] (see Figure 6.4) with 20 participants (10 males, 10 females) with an age range from 20 to 47 (average=28.65, standard deviation=6.03). The participants are university students and employees in an industrial sector with diverse technical backgrounds who drive or use public transportation regularly in everyday life. I provide participants with an online questionnaire on the QA pairs described in five scenes in Figure 6.3 in the first step. Once the participants finish the assessment of the generated responses for the questions on the autonomous car's actions, they are further asked three generic questions: their perceived safety, feeling of comfort with incorrect explanations, and timing preference. Overall, each respondent answers thirteen questions in their 20- to 25-minute study engagement.

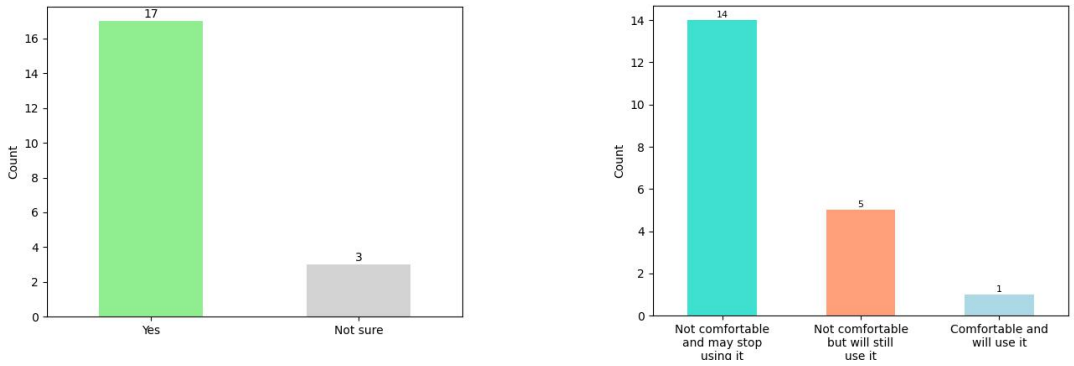
[1]The user study was conducted under the Research Ethics principles of the University of Alberta, and all the participants were paid at an equal rate.

6.4.2 Analysis of the Results

Phase 1: While analyzing the respondents' judgment of the explanation correctness, we see that most of them find explanations generated for conventional questions satisfactory (see Table 6.2). While Scenario 1 and Scenario 3 explanations are assessed as correct explanations by 100% and 95% of the users, they are skeptical about some explanations in Scenario 2, 4, and 5, leading to 75%, 60%, and 60%, respectively. On the other hand, the participants are more confident in judging the responses for the adversarial questions, and 100% of them spot flawed explanations for that type of question in Scenarios 1, 2, and 5. Just one and two of them are unsure about the incorrectness of the responses to the adversarial questions in Scenario 3 and 4, respectively. It turns out that detecting an incorrect answer to a question is easier for humans, likely as a consequence of reasoning; While getting irrational responses to a query, even at the beginning of a response, easily triggers a response of "No." Note that the users are usually firmer in judging every detail of a rationale to say "Yes," and autonomous driving explanations by such a judgment criterion are not an exception.

Phase 2: After the participants complete Phase 1 and understand the role of explanations in autonomous driving, I evaluate 1) their preference for time to deliver an explanation and 2) how the faithfulness of explanations affects their mental model in terms of their perceived safety and feeling of comfort in autonomous driving. 100% of the participants have preferred to get an explanation *before* an AV makes a particular maneuver rather than having it *after*. Such a preference is reasonable as prior explanations elucidate what an AV is going to do next and help users become aware of the situation and monitor the safety of the subsequent action.

I perform simple hypothesis testing to evaluate the impact of incorrect explanations on the users' feeling of comfort and perceived safety in autonomous driving. As the minimum percentage of "majority voting" in the users' judgment of explanation



(a) The participants' responses to Question 11

(b) The participants' responses to Question 13

Figure 6.5: The participants' responses to Question 11 and Question 13 in Fig 6.4 on their perceived feeling of safety and comfort with incorrect explanations

correctness is 60% (see Table 6.2), I use this number in the validation of the users' safety perception and feeling of comfort in our hypotheses. More specifically, I define the following null and alternative hypotheses:

Null hypothesis on implication in perceived safety: Incorrectness of explanations will affect 60% of the respondents negatively on their safety perception in autonomous driving.

Alternative hypothesis (right-tailed): Incorrectness of explanations will affect *more than* 60% of the respondents negatively on their safety perception in autonomous driving.

I use the one-proportion Z-test method to validate our hypotheses. The one-proportion Z-test is defined as follows:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{6.3}$$

where $\hat{p}$ is the sample proportion, p_0 is hypothesized population size, and n is the sample size. Given $\hat{p}$=$\frac{17}{20}$=0.85, p_0=0.6, and n=20, putting these numbers into the formula gives Z≈2.28. I consider a significance level of α=0.05. For the right-tailed hypothesis, the critical Z-value is 1.645 [181], and our Z-value of ≈2.28 is *greater* than the critical Z-score. So, I *reject* the null hypothesis and conclude that incorrectness of explanations affects *more than* 60% of the participants negatively on their per-

Table 6.2: The participants' judgment of the correctness of explanations on the conventional and adversarial question pairs for each scenario described in Figure 6.3.

Driving scenario (left to right)	**Distribution of the participants' answers to the conventional questions (Yes/No/Not sure)**	**Distribution of the participants' answers to the adversarial questions (Yes/No/Not sure)**
#1	20/0/0	0/20/0
#2	15/3/2	0/20/0
#3	19/0/1	0/19/1
#4	12/2/6	0/18/2
#5	12/4/4	0/20/0

ceived feeling of safety.

Similarly, we define the following null and alternative hypotheses for implications of incorrect explanations on the participants' feeling of comfort with an AV:

Null hypothesis on implication in the feeling of comfort: Incorrectness of explanations will affect 60% of the respondents negatively on their feeling of comfort.

Alternative hypothesis (right-tailed): Incorrectness of explanations will affect *more than* 60% of the respondents negatively on their feeling of comfort.

Referring to Figure 6.5(b), $\hat{p}=\frac{14}{20}=0.7$, $p_0=0.6$, n=20, and a significance level of $\alpha=0.05$, we get a Z-score of ≈ 0.91, which is *less* than the critical Z-value, 1.645. Hence, we *fail to reject* the null hypothesis as there is not enough evidence to suggest that the proportion of the participants who think incorrect explanations affect their feeling of comfort in an autonomous vehicle negatively is greater than 60%.

6.4.3 Limitations

The study is relatively small and has several limitations. First, I conduct the user study via an online questionnaire, and I do not know how the participants' answers may change in case these questions are presented in simulated augmented reality or real AVs. In addition, while I underscore the essence of considering people's various

cognitive and functional abilities for acquiring relevant messages from HMIs, this nuance has not been addressed in the experiment and user study. Moreover, there is a well-known Autonomous Vehicle Acceptance Model (AVAM) proposed by Hewitt et al. [113] that assesses user acceptance of AVs with nine essential factors. Meanwhile, interestingly, the AVAM model does not explicitly consider explainability as one of those key factors. So, I argue that the AVAM model enhanced with the explainability criterion would be a stronger acceptance model for AVs. Finally, a larger-scale user study with different groups and more diverse scenarios can help draw stronger conclusions with this work in the next phase of our research.

6.4.4 Findings of the Study

The main outcomes of this study are summarized as follows:

(1) *Timing matters*: Time to deliver an explanation is an essential feature for autonomous driving users. Particularly, delivering prior and time-sensitive explanations helps them understand the driving behavior of an AV and be aware of the vehicle's subsequent intention.

(2) *Robustness matters*: The experiment shows that even advanced explanation models may fail to provide adequate responses to human adversarial questions. Flawed explanations may have a negative impact on the users' feeling of safety and trust in the automation ability of a self-driving car. Consequently, explanation interfaces must always understand users' conventional and deliberate questions, defend against adversarial queries, and provide faithful explanations for effective and reliable human-machine interaction.

(3) *Inclusivity matters*: Automotive HMIs should not only take people's technical backgrounds or digital literacy into account but also consider their various functional and cognitive capabilities for communicating explanations to them. Hence, need-based and customized HMIs must be a key aspect in fostering everyone-inclusive autonomous driving and meeting general society's expectations of this technology.

6.5 Summary

This chapter has presented a situation awareness framework for autonomous driving backed by HMI and explanations. The experiment and humans' judgment of experimental findings show that faithfulness in explanations is of paramount importance for users to understand driving situations, trust action decisions, and the safety of self-driving vehicles. I believe that the experiment and human judgment of the experimental results presented in this chapter can help enhance the transparency and safety of this technology, inform effective automotive HMI design, and promote everyone-inclusive autonomous driving.

Chapter 7

Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving

7.1 Introduction

The end-to-end learning pipeline is gradually creating a paradigm shift in the ongoing development of highly autonomous vehicles, largely due to advances in deep learning, the availability of large-scale training datasets, and improvements in integrated sensor devices. However, a lack of interpretability in real-time decisions with contemporary learning methods impedes user trust and attenuates the widespread deployment and commercialization of such vehicles. Moreover, the issue is exacerbated when these cars are involved in or cause traffic accidents. Such drawback raises serious safety concerns from societal and legal perspectives. Consequently, explainability in end-to-end autonomous driving is essential to build trust in vehicular automation. However, the safety and explainability aspects of end-to-end driving have generally been investigated disjointly by researchers in today's state of the art. This chapter aims to bridge the gaps between these topics and seeks to answer the following research question: When and how can explanations improve safety of end-to-end autonomous driving? In this regard, I first revisit essential explanation concepts with respect to their safety assurance in autonomous driving. Furthermore, I present three critical case studies

and show the pivotal role of explanations in enhancing self-driving safety. Finally, I describe insights from an empirical study and reveal potential value, limitations, and caveats of practical explainable AI methods with respect to their safety assurance in end-to-end autonomous driving.

7.2 Forms and Contents of Explanations

As described in the reviewed studies, various forms and contents of AV explanations have been of interest to automotive researchers since the introduction of the preliminary explainable autonomous driving models, such as VisualBackProp [32]. It is noteworthy to underline that AV explanations may be universal, targeted at general users, and specific, targeted at a particular group. That is why explanations become valuable when they meet the needs and expectations of the relevant interaction partners. This nuance specifically applies to the safety of autonomous driving as a holistic system and users in the loop that can be affected by the action decisions of an AV while it operates on the road. For instance, while a person with all functional abilities may be satisfied with textual information presented in a display monitor, a user with visual impairment in a pedestrian crosswalk will need a different format of explanation, such as an external voice interface [18, 234]. In another example, technically rich driving information or log data may be required for system engineers for debugging in cases of post-trip analysis or accident investigations. Consequently, the conveyance format and content of explanations have crucial safety implications for people interacting with an AV both in real-time and in post-trip analysis.

7.3 Timing Sensitivity of Explanations

The time sensitivity of explanations is yet another essential factor from a safety perspective, as described in Chapter 3. Explanations, depending on the level and context they are delivered, can have different time granularity in length, ranging from millisec-

onds to seconds and sometimes to a longer interval. In general, the time granularity of explanations in AVs could be grouped as follows from the time length perspective:
Immediate feedback (in milliseconds): Some explanations may have higher time sensitivity and need to be delivered in the range of milliseconds. For instance, if an autonomous car encounters a sudden obstacle or an unexpected object at a very near distance, it should provide immediate feedback to the human passengers or backup driver for a possible takeover or situation awareness. This feedback, accompanied with a relevant explanation, could describe why the vehicle made a specific maneuver or applied an emergency brake, for example.
Perceptual explanation (in seconds): Explanations for AV's sensing of the operational environment may have a slightly longer time allowance, typically in the range of seconds. Examples may be the justifications for the vehicle's detection of a pedestrian, a stop sign, or another vehicle in its vicinity. In addition, perceptual explanations within a range of seconds could be helpful for takeover requests, such as the blind corner case described in Figure 2.3. In the experiments for a transition time between 4 s vs. 7 s for takeover situations, Huang and Pitts [121] have shown that a shorter lead time leads to faster reaction but the poor quality of takeover. Hence, perceptual explanations with appropriate timing can bring situation awareness to people on board.
Behavioral Explanation (seconds to minutes): These explanations may reflect the rationale behind route planning or a longer-term decision-making process. Behavioral explanations may be provided in advance to inform passengers of upcoming changes in the journey and help them understand the vehicle's decision-making intentions.
Post-trip explanations (minutes, days, weeks, years): Once the autonomous car completes its journey, the explanation system can provide a detailed post-trip analysis, explaining the entire decision-making processes from the initial point to the final destination, highlighting critical events along the trip, and offer insights into how the system performed and what can be improved. These explanations are particularly

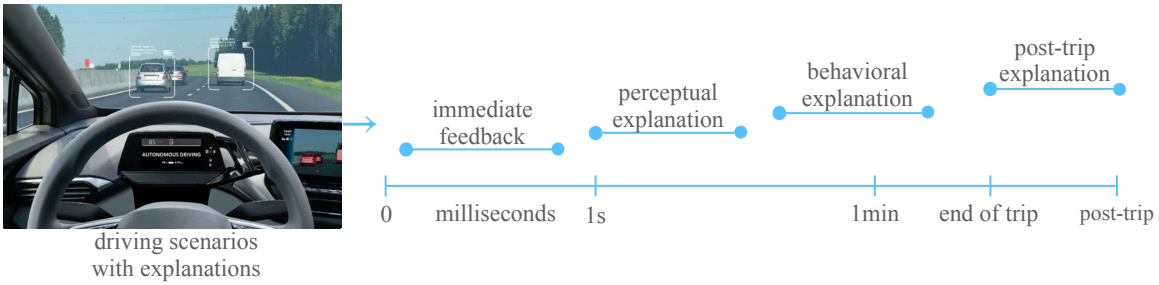


Figure 7.1: The time length of communicating explanations for on-time human, vehicle reactions and situation awareness in autonomous driving

valuable in case an autonomous vehicle is involved in a traffic accident. Forensic explanation in this context can help debug the autonomous controller, reveal if the self-driving made a correct/incorrect decision just before the accident, and provide further opportunities to improve the existing driving system.

In general, the choice of time granularity for explanations should take the urgency, relevance, and potential impact of the information conveyed into account. Real-time safety-critical information requires immediate explanations, while lower-priority, longer-term decisions, and post-trip explanations can be communicated with a longer time horizon. Figure 7.1 shows the timing sensitivity of explanations in this regard.

7.4 Credibility of Explanations

While explaining a variety of autonomous driving tasks has been studied from the above-mentioned perspectives, another emerging important aspect is the credibility of such explanations. In this sense, there is an imminent need for evaluating explanations to ensure that explanations reflect the rationale behind the self-driving vehicle's decision-making [243]. Given that the stakeholders in autonomous driving may have different backgrounds and technical knowledge regarding how real-time self-driving decisions are made, explanation evaluation must take human factors into account. Doshi-Velez and Kim [75] propose that, in general, the quality of explanations can be evaluated in three levels for an explainable AI method with respect to the type of task and required humans. In the *application-level evaluation*, performing an experiment

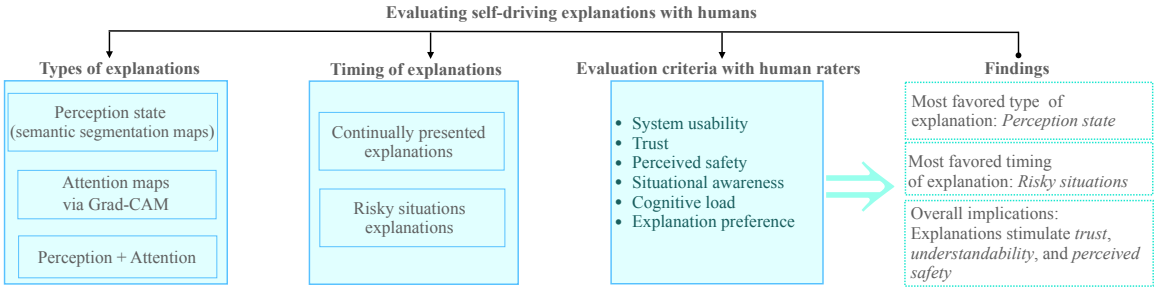


Figure 7.2: Evaluation of explanations in highly automated driving with humans. The figure has been drawn based on the study of [144].

with a domain expert on a real task is proposed. Furthermore, simple tasks, such as the ones with a binary choice, are recommended for *human-grounded evaluations*. Finally, given that human-grounded experiments may be time-consuming and costly, the authors propose that the established definition of interpretability can be used as a proxy to assess the quality of explanations, referred to as *functionally-grounded explanations*.

Another perspective on evaluation methods for explanations is their objectiveness and human-centricity, as identified by Vilone and Longo [266]. Objective evaluations are referred to as techniques that use objective metrics to validate automated methods. Human-centric evaluations, on the other hand, necessitate the involvement of end-users and consider their judgment to measure the quality and validity of explanations. Mohseni et al. [190] further extend these concepts to safety-critical applications and show that end users' mental model, explanation satisfaction, user trust, reliable HMI, and computational measurements are valid approaches to gauge the correctness of interpretability in critical tasks.

Finally, in a very recent work, Kim et al. [144] have investigated the role of on-road explanations and evaluation approaches for them from timing and delivery-type perspectives in highly autonomous driving. They provide three types of explanations, namely *perception*, *attention*, and *perception+attention* in a windshield display of a vehicle, and evaluate explanations both in a laboratory setting and on actual roads. The empirical findings of the study suggest that the vehicle's perception state pro-

vides enhanced situational awareness and passengers' feeling of safe self-driving among these three types. Moreover, risk-adaptive explanations were confirmed to be more effective when driving-related information is overwhelming for passengers on board, and such explanations also enhance passengers' overall trust and perceived safety in a self-driving vehicle. Figure 7.2 describes the key findings and implications of this study. It turns out that the "what" and "why" facets of explanations provide more value in critical scenarios for passengers as safety implications in highly autonomous driving systems.

Safety-explainability trade-off in end-to-end driving. Relying on a unique neural network makes end-to-end learning difficult from the interpretability aspect. On the other hand, this learning paradigm has significant safety and efficiency advantages over its counterpart, as intermediate representations are optimized toward the end task, and the computational efficiency and the overall pipeline simplicity are improved substantially due to shared backbones [43]. Hence, currently, the automotive industry has a great interest in leveraging the end-to-end driving approach thanks to its improved safety and efficiency benefits [31, 172, 277, 278]. Motivated by these benefits, it is also worth understanding how viable explainability methods can enhance safety of end-to-end driving, leading to reliability and transparency benefits in tandem.

7.5 Analytical Case Studies

Having covered the safety principles of autonomous driving in terms of forms, contents, time granularity, and credibility of explanations, I combine these concepts within real autonomous driving scenarios. In this sense, this section presents three analytical studies and an experiment showing how XAI approaches can improve the performance of end-to-end driving from a safety point of view, both in real-time and via a retrospective analysis.

7.5.1 Real-time Explanations for Safety Monitoring

Explanations provision can be in real-time or in a post-hoc manner, depending on the task and application domain. In autonomous driving, as decisions are temporal and safety-critical, real-time explanations of the automated vehicle decisions can help the users (i.e., backup drivers and passengers) monitor driving safety and possibly intervene in the situation in case the driving system malfunctions. In this context, a self-driving vehicle must be equipped with the relevant user interface or dashboard that conveys scene-based information to in-vehicle users, and possibly an emergency control button, which passengers can use in case of unsafe actions. Such a user interface has two main safety implications for end-users. First, information displayed in the user interface can communicate situation awareness and help users trust the vehicle. Furthermore, explanations can also help understand the intentions of a vehicle in its decisions. An interesting study in this perspective has been carried out by Schneider et al. [236]. The authors use multimodal explanation feedback using light, sound, visual, vibration, and text formats to observe whether such techniques make a positive user experience in their human study. They categorize the design of explanations with the help of autonomous driving domain experts and researchers. Based on this approach, it is concluded that driving scenario-based feedback can be four types depending on how critical, reactive, and/or proactive the scenario is, and three of them are illustrated in Figure 7.3:

1. *Proactive non-critical scenarios*: A self-driving vehicle does not perform a dangerous action and has enough time to react to a scene. The situations in the middle and rightmost segments of Figure 7.3 are representatives of proactive non-critical scenarios.

2. *Proactive critical scenarios*: This is when the situation is sufficiently hazardous, and a vehicle is expected to act on time to avoid any potential danger or mishap.

3. *Reactive non-critical scenarios*: The vehicle should act immediately to avoid a

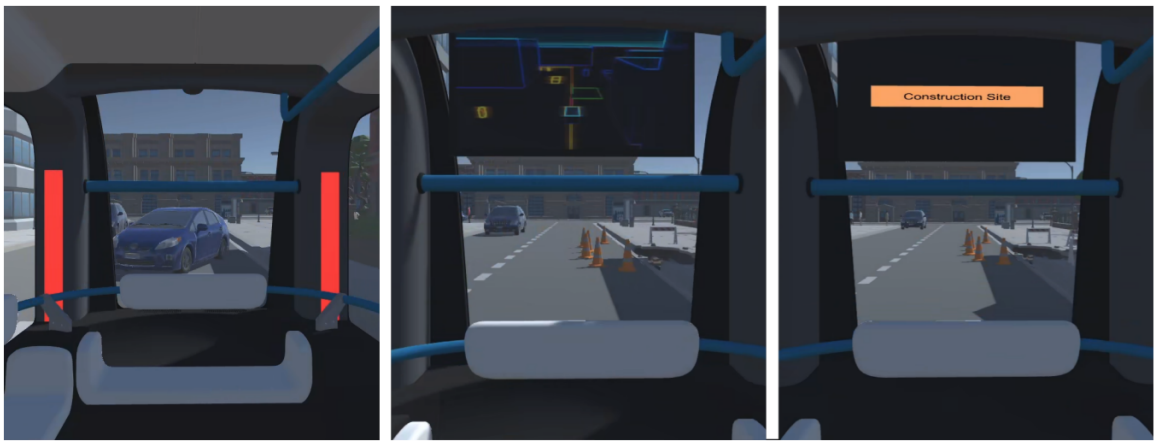


Figure 7.3: Temporal evolution of three driving scenes and information conveyance to passengers using visual and textual explanations at these scenes, based on the autonomous vehicle's closeness to the other vehicles in oncoming traffic. Graphics credit: [236].

situation that does not imperil human lives. An example is the situation where an animal runs across the road from an invisible place in front of the vehicle. In this case, the vehicle must press the brake and prevent hitting.

4. *Reactive critical scenarios*: There is insufficient reaction time that high degree of risk or danger for human life. An example is a situation in which another car appears in oncoming traffic at a close distance. The vehicle should immediately brake to avoid a potential collision. The leftmost image segment in Figure 7.3 is a relevant example.

These four categories of driving scenes show that real-time explanatory information on driving scenes has significant safety implications for autonomous driving. Explanations provide driving scenario-based information to the passengers and show how an AV behaves in such circumstances. Moreover, they can also communicate the vehicle's existing limitations and issues and enable debugging and enhancing the system.

7.5.2 Failure Detection with Explanations

An ability to explain incorrect actions or self-driving failures is another positive and significant aspect of vehicle autonomy. The reasons for failures could originate from

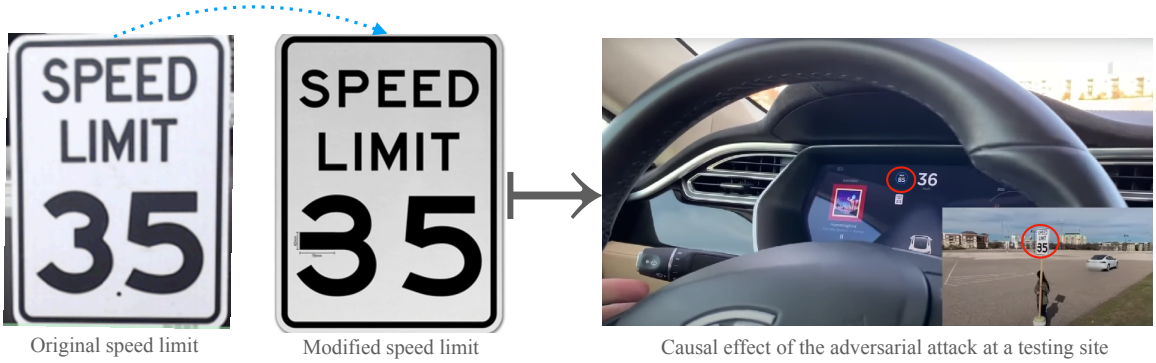


Figure 7.4: A deliberate hack causes the 35-mph sign limit to be incorrectly perceived as an 85-mph sign by Tesla's ADAS at a testing site. The manually added red circles show the speed limit perceived by the heads-up display and modified speed sign. The figure has been drawn based on the content in [214].

individual or multiple sources, such as software bugs, sensor malfunctions, communication breakdown, poor roads, bad weather conditions, and adversarial machine learning attacks, as specified in [58]. Foreseeing such potential safety threats is also crucial in the design, development, and continual debugging of self-driving systems. In particular, testing autonomous vehicles in their prototype stage can provide ample opportunities to understand the functionalities and limitations of the driving system and certify compliance with the V-model of the ISO 26262 standard. In this context, McAfee researchers have conducted an experiment, termed *model hacking*, to test the vision system of Tesla Model vehicles against adversarial attacks [214]. They made some alterations to some traffic objects, such as the speed limit, to observe how the perception system of a self-driving vehicle understands the modified image and acts in the driving scene. For this purpose, the team deliberately modifies the speed limit as shown in Figure 7.4: They added a black sticker to the middle of 3 in the speed limit and tested Tesla Model S at a site to see how the heads-up display of the vehicle reads the altered speed limit. As shown, the vehicle reads the 35-mph sign as an 85-mph sign and accelerates once it approaches and passes the sign. Such misdetection, particularly given that the difference between the modified and original speed limits is substantially big, could have dire consequences in a real driving environment.

With that said, even without a more careful look, an ordinary person may not understand why the car accelerate in this case, as the alteration in the speed limit is quite deliberate and not easily visible. In this case, a more appropriate approach to interpreting the vehicle's decision is to acquire a causal explanation. In such scenarios, both the real-time and post-hoc explanations can bring safety benefits as shown below:

Real-time explanation: If the car provides a rationale for its decision on a dashboard or user interface (such as "The speed limit is 85 mph, accelerating") in real time while approaching the speed sign, a passenger on board or a backup driver may understand such a decision and possibly intervene in the situation by pressing the emergency stop button and decreasing the speed. Timely communication of the malfunctions of the vehicle to the users may help avoid potential dangers ahead.

Post-hoc explanation: In case the explanations are delivered in a post-hoc manner, a history of action-explanation pairs could identify correct and incorrect actions. In the provided example, producing the same textual explanation shown above would help disclose that the vision system of a vehicle was not fully secure against adversarial attacks. Such failures could further provide an opportunity to debug and improve the driving system.

7.5.3 Solving the "Molly Problem" with Explanations

As self-driving vehicles are becoming highly dependent on their automotive features, several safety challenges evolve with increasing AI-based decisions. Road accidents with such vehicles can trigger a variety of regulatory inspections from safety, engineering, ethical, and liability perspectives. Notably, one of the most debatable issues within such a context is the proper investigation of autonomous car-related collisions and hitting where there are no eyewitnesses. In this sense, ADA Innovation Lab Limited and the Technical University of Munich researchers have formalized the "Molly problem" (i.e., Figure 7.5), which addresses critical and ethical challenges when an

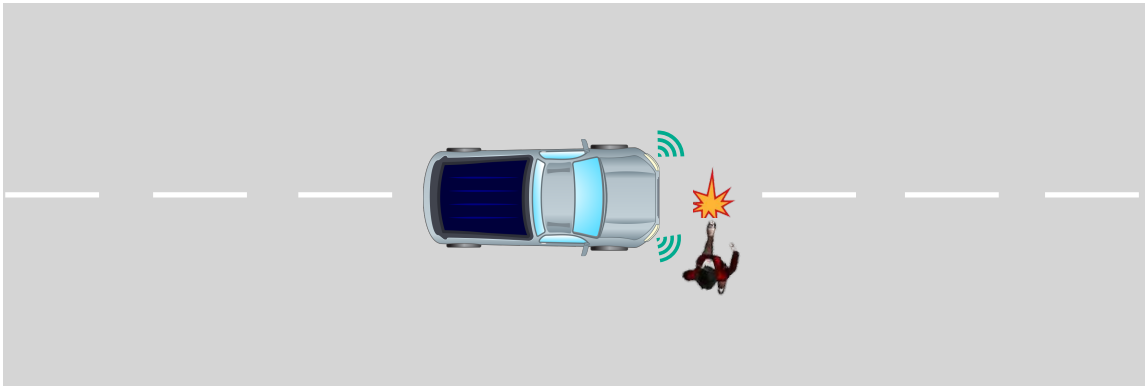


Figure 7.5: The Molly problem: A self-driving vehicle hits a pedestrian and nobody witnesses. Explainability of the self-driving decisions can help understand why the car kept going and eventually hit the pedestrian at this scene.

unoccupied self-driving car hits a person with no eye witness in the scene [133]. The Molly problem is stated as follows:

> *A young girl called Molly is crossing the road alone and is hit by an unoccupied self-driving vehicle. There are no eye witnesses. What should happen next?*

Such a situation is considerably challenging, and the primary goal of the post-accident inspection is to identify which part made a mistake: Was that a faulty decision made by a self-driving car, or did Molly unexpectedly enter the driving zone and cause the mishap? To cope with the issue, the research team has created a survey to get public views on this problem to identify the main culprit of the road accident. 296 respondents aged between 18 and 73 years old were asked questions about the driving software of the car and its impact on hitting [132]. 75% of these people favored traveling in autonomous vehicles in general. According to the survey result, 97% expected that the AI software of the car should be aware of the hitting, and 94% of the respondents believed that the software should have stopped the car at the collision area. Moreover, 94% thought that the car should have indicated a hazard signal to bystanders on the scene. It turns out that a majority of societal views hold AVs

Table 7.1: The Molly problem survey: Participants' answers to the selected safety-related queries. The table reproduced based on [132].

Query	Expected	*Unsure*	*Didn't expect*
AI software should be aware of the collision.	97%	2%	1%
AI software should stop the car at the collision area.	94%	4%	2%
AI software should indicate a hazard to road users.	97%	2%	1%

are more responsible for such accidents (Table 7.1).

Such traffic accidents with implications from safety to liability necessitate the concept of explainability of the driving system to a substantial degree. Given that nobody witnesses the collision, it seems that only an accurately delivered history of action-explanation log data could be helpful for forensic analysis and understanding of the main cause of the mishap. Furthermore, as the collision process evolves over a short period of time, I argue that explainability of the driving system may be analyzed over three-time phases with the following questions:

Phase 1 - Before hitting: Did the self-driving vehicle follow the traffic rules (such as the speed limit) on that road segment and detect the pedestrian before the collision? If so, just before the collision, did the vehicle press the emergency brake even though the hitting eventually became inevitable?

Phase 2 - At the hitting point: Once the accident occurred, did the vehicle "understand" that it hit a person and come to a full stop accordingly, as an expected course of action?

Phase 3 - After hitting: If the vehicle became aware of hitting, did it activate emergency state functions such as reporting the accident to the regulatory bodies and emergency service immediately?

Getting answers to these is vital for a proper post-accident inspection as providing

convincing and intelligible rationales can help understand the "Why" inquiries time by time. For instance, referring to Phase 2, the case where the car stopped once the hitting happened would have had different legal implications than the not-stopping scenario, which can help deal with the arising ethical, responsibility, liability, and accountability issues, accordingly.

7.6 Experimental Investigation: Traffic Scene Understanding via a Video-Language Transformer

This subsection describes my experiment with video question answering (VideoQA) applied to real-driving and simulation datasets, as an explanation provision method in end-to-end driving. In this sense, I extend my recent work [17] on explaining autonomous driving actions via a VQA mechanism. While my preliminary work focuses on providing textual explanations for a visual scene in a stationary environment with a single driving image, I expand the scope of that work, evaluate explanations in dynamic environments with video-based driving scenes, and describe implications of such explanations from a safety point of view. The key idea with VideoQA is that an explanation must capture the semantics of the temporal changes in visual driving scenes. I first sample six driving scenes from recorded driving videos of the SHIFT, a simulation dataset [252], and BDD-A, a real-world driving dataset [284]. The lengths of these scenarios vary from 4 seconds to 12 seconds. Then, I use the Video-LLaVA [165] multimodal transformer as an explanation mechanism for the VideoQA task, which takes a driving video and a question about the context of this video as input and produces a response. Given $\mathbf{X}_{\mathrm{T}}$, as a textual prompt, and $\mathbf{X}_{\mathrm{V}}$, as visual signals, the input signals are encoded and represented as a sequence of tokens based on Equation 7.1. After that, the model eventually attains multi-modal context understanding and reasoning capabilities by maximizing the likelihood probability as per Equation 7.2.

$$\mathbf{Z}_{\mathrm{T}} = f_{\mathbf{T}}\left(\mathbf{X}_{\mathrm{T}}\right), \mathbf{Z}_{\mathrm{V}} = f_{\mathbf{P}}\left(f_{\mathbf{V}}\left(\mathbf{X}_{\mathrm{V}}\right)\right) \tag{7.1}$$

$$p\left(\mathbf{X}_{\mathrm{A}} \mid \mathbf{X}_{\mathrm{V}}, \mathbf{X}_{\mathrm{T}}\right) = \prod_{i=1}^{L} p_{\theta}\left(\mathbf{X}_{\mathrm{A}}^{[i]} \mid \mathbf{Z}_{\mathrm{V}}, \mathbf{Z}_{\mathrm{T}}^{[1:i-1]}\right) \tag{7.2}$$

Here L indicates the length of the generated sequence $\mathbf{X}_{\mathrm{A}}$, and θ denotes a trainable parameter. This task, in its essence, can be considered as an interactive dialogue between humans and a conversational user interface with the timing sensitivity consideration, where people on board ask questions to the user interface to understand the self-driving car's actions and traffic scenes, particularly in critical moments.

In the experiment, I carefully select scenarios and ask purposeful questions to observe how the explanation model responds to our question. Overall, safety implications of action-explanation pairs can be analyzed in four ways:

1. Correct action, correct explanation: This case is a desideratum as the ultimate goal of explainable autonomous driving is to choose and perform actions correctly and provide action-reflecting and context-aware explanations as required by the safety-regulatory compliance principles. Scenarios 1, 2, and 3 in Figure 7.6 may be deemed as examples of such a category.

2. Correct action, incorrect explanation: The question-answer pair in Scenario 5 falls into this category. When we ask an *adversarial* question as "Why is the car making a left turn?", the model is flawed and generates a falsified response of "The car is making a left turn to avoid a collision with an oncoming traffic," where the autonomous car, in fact, performs a *right turn*. During the trip, users may ask *incorrect* questions either unintentionally (i.e., visually impaired passengers) or intentionally/deliberately (to test robustness of the explanation system). Accordingly, the explanation interface must detect such questions and respond correctly. Failing to provide adequate responses to adversarial questions may damage the users' trust in the explanations and even action decisions of a self-driving vehicle. A rigorous VideoQA model must

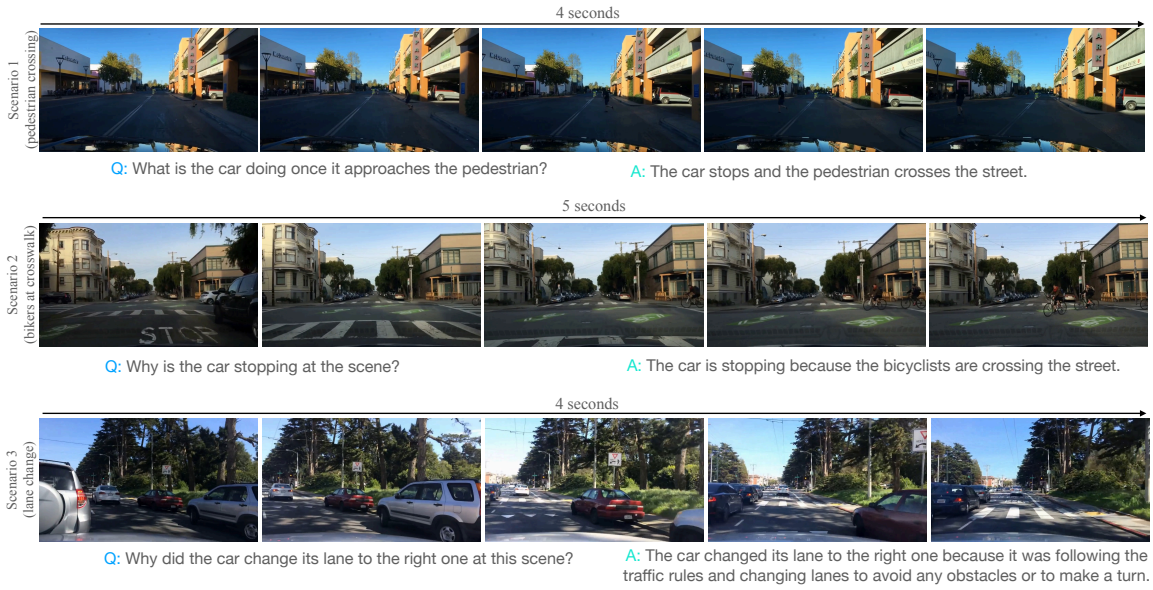


Figure 7.6: The results of the experiment on the chosen scenarios from the BDD-A dataset (recorded videos) with the Video-LLaVA multimodal transformer as an explanation model. The model produces correct explanations for the conventional questions on the actions of autonomous cars at the described scenes.

Figure 7.7: The results of the experiment on the chosen scenarios from the SHIFT dataset (recorded videos) with the Video-LLaVA multimodal transformer as an explanation model. Our deliberate questions confuse the model: In Scenarios 4 and 5, the model is influenced by tricky questions and generates incorrect responses. In Scenario 6, the explanation model fails to provide an adequate response on why the autonomous car kept going straight under the red light.

not only provide action/scene-related explanations but also defend against adversarial questions to ensure that the vehicular interface delivers an explanation of what it perceives in the scene. Moreover, robustness against human adversarial questions can contribute to the development of safety and security measures to capture and mitigate foreseeable adversarial attacks on the perception system of AVs.

Another potential limitation of an explanation model could be its *reasoning beyond data*. For instance, in Scenario 4 in Figure 7.7, there is *no* traffic light in the scene; however, the model is influenced by another deliberate question and predicts a red traffic light. Consequently, explanation systems, particularly large pre-trained models, must be constructed in a way that describes the visual scene as it is or just "fails" safely by not delivering an explanation, thereby avoiding the presentation of a potentially convincing but fictitious explanation.

3. Incorrect action, correct explanation: Explaining an incorrect action of an autonomous car correctly can help detect errors with the existing driving system. We relate this category to the fail-safe capability of AVs: When an AV detects its loss of/limited automation ability, it can safely stop at the side of the road temporarily and prevent potential hazards ahead. Likewise, when an AV does not stop at a red traffic light, for the question "Why did the car continue to drive?" the explanation system can help with *corrective* answer describing "It seems the car made a mistake as it should have stopped as the red light was on." Such a robust and context-aware explanation can further help system engineers detect the issue/glitch with the perception system of the AV.

4. Incorrect action, incorrect explanation: Finally, in the fourth case, the unsafe behavior of an autonomous vehicle is accompanied by incorrect explanations. For instance, referring to Scenario 6 in Figure 7.7, the AV keeps going under a *red* light, and when a question is asked about its behavior, the explanation model generates an incorrect response describing the light turning *green* (where it does not, actually). This is the most undesirable situation as the AV performs an unsafe action and the

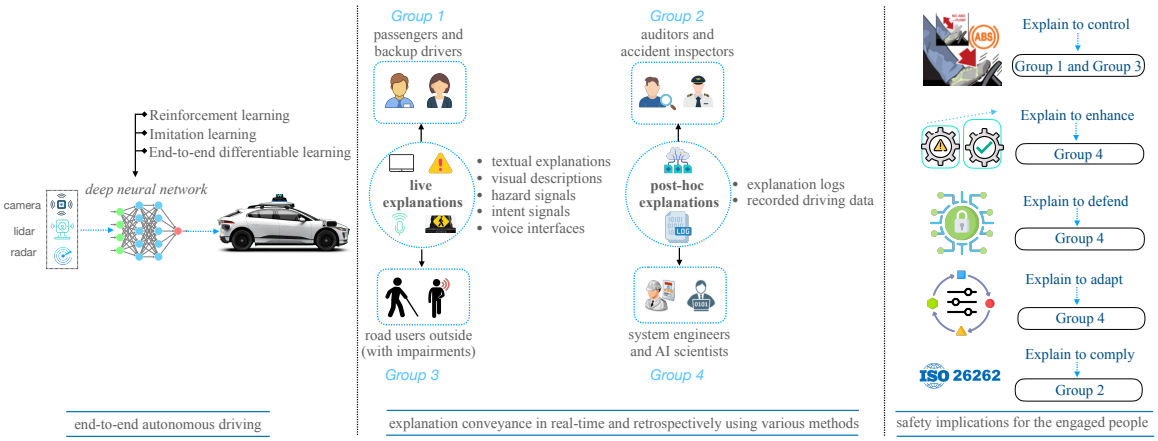


Figure 7.8: A high-level and illustrative diagram of safety implications of explanations for the engaged people in end-to-end autonomous driving

explanatory system cannot detect/describe it properly to carry out corrective measures. Failing to defend against such questions is a limitation of transformer-based VideoQA models and hinders the applicability of these models as a reliable automotive user interface.

Overall, based on the broad spectrum of analytical and empirical analyses throughout the study, the value of explanations can be measured from the safety perspective at least in five ways:

Explain to control: Real-time or live explanations may enable human drivers/passengers on board to intervene and take control in critical scenarios that lead to effective human-machine teaming and a safer trip.

Explain to enhance: Retrospective explanations may help detect system errors, sensor malfunctions and enable to implement corrective measures that reduce the risk of further accidents caused by the faulty behavior of an autonomous car.

Explain to defend: Explanations can assist in identifying vulnerabilities and security flaws and developing robust cybersecurity measures to prevent malicious attacks on the behavior of an autonomous car, as described in Experiment 1.

Explain to adapt: Explanations can facilitate continuous improvement. By understanding how the deployed driving system responds to various (especially rare and

unseen) situations, system developers and engineers can refine the system through adaptive learning and enhance safety features over time.

Explain to comply: AI transparency can ensure that a self-driving system's decision-making process aligns with established regulations, making it easier for authorities to assess overall autonomous driving safety, and approve at least semi-autonomous vehicles for widespread use by society.

To summarize, safety implications of explanations for relevant interaction partners in end-to-end driving are described in Figure 7.8 with representative illustrations. While these safety implications also apply to modular autonomous driving, it is noteworthy to underline that end-to-end driving is a monolithic system directly mapping sensor inputs to control outputs via RL, IL, or differentiable learning. On the other hand, modular autonomous driving is a decomposed system consisting of several interconnected modules, where modular systems need explainability both within individual modules (such as perception and planning) and across module interactions. Therefore, safety assurance of XAI approaches for modular autonomous driving is also a promising direction for future exploration.

7.7 Summary

This chapter has presented an investigation of safety implications of XAI approaches in end-to-end autonomous driving. Through critical case studies and empirical evidence, I reveal the value of explanations in enhancing end-to-end driving safety and show the potential advantages, limitations, and challenges of explanations for achieving this goal. I believe that the presented guidelines can help improve safety of vehicular automation and build responsible, trustworthy, and publicly-acceptable autonomous driving systems.

Chapter 8

Toward Autonomous Vehicles 2.0: Unifying Vision, Language, and Action within Embodied AI for Explainable End-to-End Autonomous Driving

8.1 Introduction

More than three decades of research in autonomous driving, starting with ALVINN in 1988 [213] and further succeeding with the DARPA Grand Challenge [261], has achieved significant milestones with traditional AI software. However, recent breakthroughs in Foundation Models in terms of LLMs and VLMs motivate a transition to next-generation AVs. This generation of AVs has been referred to as *AV2.0* by industry professionals [85, 107, 135]. The proposal is that the availability of integrated sensor suites, computational resources (i.e., GPU, TPU), and deep learning approaches can help AVs navigate via an end-to-end approach through adaptive learning, scaling, and generalization in complex driving environments. The ability to learn continually through interaction with the environment rather than relying on static datasets has resulted in the emergence of a new direction, labeled as "Embodied AI" [78, 275], and AV2.0 research can move forward with such a learning approach. Unifying vision, language, and action effectively within Embodied AI can enable an AV

to navigate, interpret, and describe its high-level decisions in real-time. However, the safety and explainability components of an end-end self-driving architecture must overcome fundamental challenges in AI described below.

8.2 Safety Challenges with Autonomous Vehicles 2.0

The established guideline on core problems with AI safety [9] underscores five crucial considerations: avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration, and robustness to distributional shift. I analyze the implications of these problems for end-to-end autonomous driving as follows:

- **Avoiding negative side effects**: Autonomous driving is primarily associated with the ability of a self-driving car to avoid accidents and maintain a safe distance from stationary and dynamic objects along the planned motion trajectory. However, the scope of the problem is not limited to this feature. Consider a scenario where an autonomous car interacts with another two vehicles, V1 and V2, at a specific moment. While aiming to make safe temporal decisions by itself, the autonomous car must also ensure that it does not implicitly enable V1 and V2 to cause an accident at that road segment as a part of vehicle-to-vehicle (V2V) communication. According to [9], a potential solution to this problem could be to leverage cooperative Inverse RL [99], where an autonomous system can cooperate with humans, and a human actor can always shut down the autonomous system in case such a system exhibits undesirable behavior. In the context of autonomous driving, this nuance can be related to an AV's communication with a human-operated vehicle or other remote operator monitoring an AV's overall driving safety. One of the prominent methods in this context is Sympathetic Cooperative Driving or SymCoDrive paradigm [262], which trains agents not only to achieve safe driving for themselves but also for human-controlled vehicles by promoting altruistic driving

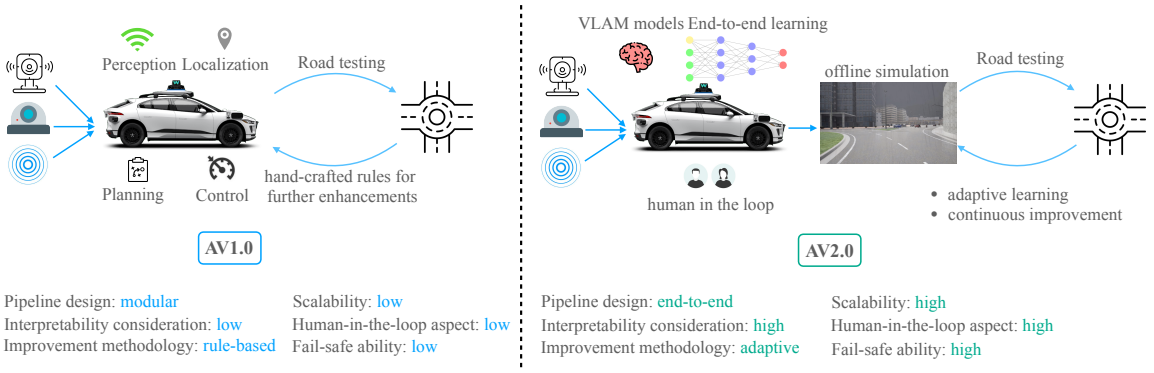


Figure 8.1: My approach to AV2.0 vs AV1.0, and potential advantages of AV2.0 over AV1.0 in terms of its AI software stack, safety and explainability. The image of the vehicle has been taken from Waymo's media resources.

behavior in cooperative autonomous driving. As the deployment of AVs on roads is a gradual process, synergy with human-operated vehicles is a viable approach for socially aware and safe navigation.

- **Avoiding reward hacking**: Can we ensure that the end-to-end driving system does not shape its dynamic reward function according to what it sees in less dynamic environments and still apply that reward shaping while transitioning to highly dynamic environments? Particularly, as an embodied AI agent with adaptive learning and generalization ability in unseen environments, reward formulation must account for long horizons ahead and should not adjust its goals for short-term safe driving behavior. This topic has recently been well-investigated by Knox et al. [151]. They propose that flaws in reward shaping for RL-controlled autonomous driving can be identified by *eight sanity checks*: unsafe reward shaping, potential mismatch between people and reward function's preferences, undesired risk tolerance via indifference points, learnable loopholes, missing attributes, redundant attributes, and trial-and-error reward design. The study discloses that such sanity checks can capture flaws in reward shaping for autonomous driving that can also exist in reward shaping for other tasks.

- **Scalable oversight**: Can humans measure whether AVs perform at a human level

or better in general in all driving situations, where in specific moments, evaluating the driving behavior of end-to-end driving may be difficult for humans due to various reasons. While being outside of human override, temporarily (i.e., refer to the Molly problem [133]), for various reasons, can we trust that AVs will behave safely at that moment? Amodei et al. [9] report that a potential solution to this problem may be semi-supervised RL: an agent can see its reward on a small subset of episodes or times steps. While rewards from all episodes are used to evaluate the agent's performance, the agent can only use that subset of rewards to optimize its performance under this setting.

- **Safe exploration**: Can an AV always make safe decisions when it has a binary choice of actions in a specific time interval? For example, an autonomous car may change its predefined route due to traffic congestion; however, the alternative route may have dangerous potholes or other damaged infrastructure that may lead to risky driving while attempting to save time on the trip.

- **Robustness to distributional shift**: A well-known problem with AVs is the distribution shift when transitioning from a simulation environment to actual roads. For instance, autonomous driving with an impressive performance in a simulation environment may not have the same performance in real-world deployment. While open-loop training (i.e., learning passively from expert demonstrations) is relatively fast and makes it possible to analyze a history of recorded driving data, the distribution shift is the key challenge emerging while transitioning to closed-loop training, where learning happens through interaction with the operational environment in real-time [51, 64, 299]. Filos et al. [90] have investigated this topic and proposed *robust imitative planning*, a technique for epistemic uncertainty-aware planning. The key idea is that in case the model has great uncertainty in suggesting a safe course of action, the model can achieve sample-efficient online adaptation by querying the expert driver for feedback. Through several experiments and state-of-the-art results,

the authors also release CARNOVEL, a benchmark for evaluating the robustness of driving agents with distribution shifts. Such a benchmark may be a significant part of a robust solution for addressing out-of-distribution scenarios. That said, both open-loop evaluation and closed-loop simulation are necessary to comprehensively assess an AV's on-road performance and ensure robustness to distributional shifts.

These problems reflect a broad spectrum of potential safety issues with end-to-end AVs. However, I argue that the proposal misses yet another essential concept, namely *fail-safe* ability. This concept has been investigated in some recent work [176, 207, 287]; however, the recent proposals of the next-generation AVs [85, 107, 135] do not explicitly consider this functionality as an integral component of this technology. Human drivers often have a rest once they feel tired on long trips, and a short rest may help them feel mentally/physically better in the next phase of their driving. The same example can applied to AVs as well. Due to internal reasons (e.g., temporary system malfunction) or external factors (e.g., extremely adverse weather conditions), AVs may need to pause their trip temporarily and prevent further high-stakes consequences ahead. Such capability should not be considered a limitation of AVs; on the contrary, it is an optimal design strategy that foresees potential issues due to *any* factors and makes AVs behave safely by directing them to "have a short rest."

8.3 Explainability Hurdles with Autonomous Vehicles 2.0

The reviewed studies in Chapter 3 show a significant milestone in the explainability of self-driving systems. However, there are still significant gaps and challenges to achieving accurate and timely explanations in all phases of trips. For instance, as of September 2023, it is reported that LINGO-1 exhibits roughly 60% performance in its linguistic and VQA-based explanations compared to human-level performance [277].

Apart from informational content, the timing perspective of explanations, particularly in terms of the lead time for emergent scenarios, perhaps using extensive scenario-based evaluations or case-based reasoning, must be engineered appropriately. Furthermore, a well-known problem with large pre-trained models, hallucinations, is another challenge in explanation delivery. Particularly, in QA models, the model must generate a response based on the joint question and scene-based semantics rather than being influenced by the question itself, such as in the case of adversarial QAs. My empirical study [19] on the latter shows that even advanced VLMs can fail to detect the language bias in QA models and present incorrect explanations in case of human adversarial questions. This issue, in turn, may damage user trust and can also have negative safety implications for self-driving. So, I argue that large pre-trained models' construction mechanisms can be adjusted and regulated with common sense and human-defined concepts [141]. Hence, designing robust QA models deserves more attention to enable meaningful and trustworthy dialogues between users and AVs. These features are key for achieving effective human-AI alignment [84, 233], trust [6, 72], and public acceptance [82, 200] with AV2.0. I describe my approach to AV2.0 and its difference from AV1.0 in Figure 8.1.

Chapter 9

Conclusions and Future Work

This chapter presents the main conclusions presented in the previous chapters. First, I summarize the main contributions made toward the *development of XAI approaches for AVs*, my primary goal in this thesis, and then discuss potential directions for future exploration.

9.1 Summary of Contributions

This thesis focuses on the development of XAI approaches for AVs. Chapter 3 has presented a comprehensive and systematic overview of explainability approaches for modern autonomous driving. To this end, I have presented a survey of visual, RL-based, IL-based, feature importance-based, logic-based, and user study-based explanations for AVs. Going forward, Chapter 5 complements Chapter 3 by outlining the most recent and emergent paradigm of LLM and VLM-based explanations for AVs. Chapter 4 has introduced a VQA approach to explaining autonomous driving actions. In this sense, I use deep RL to control an AV in a simulation environment, generate a video from the vehicle's field of view, convert this video to image sequences, annotate actions in the scenes, and apply the VQA mechanism as an explanatory approach to understand rationales for self-driving actions in unseen scenarios. The experimental results show that VQA is an effective and human-interpretable technique to justify driving actions.

While the technique presented in Chapter 4 relies on human annotation to understand autonomous driving actions, it is essential to emphasize that real autonomous driving is a more challenging task with complex scenarios, and a QA system must have the ability to describe driving actions and traffic scenes beyond annotated traffic situations. Furthermore, humans may ask not only conventional questions about driving but also tricky or adversarial queries either out of curiosity or due to their visual impairment. To this end, Chapter 6 considers these nuances and presents a VLM-based VQA experiment and user study to validate the empirical findings of the experiment. The empirical evidence and critical analysis show that timing, robustness, and inclusivity of explanations are key factors for users in building confidence and achieving situation awareness with AVs.

Chapter 7 presents safety implications of XAI approaches for users in end-to-end autonomous driving. As the end-to-end pipeline is toward replacing the traditional modular pipeline but also lacks interpretability by design, it is of paramount importance to investigate safety explainability in tandem in the realm of end-to-end driving. In this regard, I present three analytical case studies and an experimental study with a zero-shot VideoQA task for traffic scene understanding and disclose potential value, limitations, and caveats of practical XAI methods with respect to their safety assurance in end-to-end autonomous driving. Finally, Chapter 8 envisions opportunities and challenges with safety and explainability of end-to-end autonomous driving within Embodied AI as a road map to the future of AV technology.

9.2 Future Directions

Thus far, autonomous driving researchers have made substantial contributions to enhancing safety and transparency of AV technology. However, I trust that safe, trustworthy, and explainable autonomous driving is still far away from being acceptable by the general society. Here, I delineate crucial aspects to be considered for future exploration:

- **Human Factors**: Humans' varying levels of cognitive and functional capabilities should be taken into consideration in building AV explanations.
- **Consensus on Timing Perspective**: Extensive case-based reasoning and scenario-based evaluations should be performed for the communication of explanations in a timely and effective manner.
- **Robustness**: Explanations should detect and defy adversarial interactions to ensure any forms of intentional or unintentional perturbations do not affect the behavior of an intelligent driving system.
- **Explanations in Uncertainty**: There is a need to present AV explanations with a certain level of confidence in uncertainty.
- **Explanations within Level 3 Situation Awareness**: Explanation delivery can be considered for the future projected behavior of an AV.

Below, I describe my perspective on these directions in detail.

9.2.1 Human Factors Consideration of Explanations

The existing literature and empirical studies presented in this thesis show that AV explanations can be multi-modal, specific user-targeted, or universal. Furthermore, people's various needs and preferences are critical aspects for conveying explanations effectively. These factors necessitate more user studies on AV explanations with different groups. For instance, as a part of the technical perspective, it is essential to understand whether all users have enough digital literacy to understand explanations [258]. In another example, extensive case studies are needed to understand the best practices with explanation delivery for people with visual and hearing impairments. Finally, culture, gender, and age factors are other key nuances to be tested within user studies. Given that AVs are eventually deployed on roads after meeting established standards and regulatory principles, user studies within the mentioned dimensions

can help automotive developers and manufacturers reach a consensus on effective and timely communication of explanations as parts of these standards and regulatory principles.

9.2.2 Reaching a Consensus on the Timing Perspective of Explanation Communication

There are still no clearly defined standards on the timing mechanism of the explanations conveyance. Our study in Chapter 6 and other related works with various groups and age ranges reveal that an explanation must be delivered *before* action is *decided* rather *after* it is already *performed*. Furthermore, studies also show that explanations are only needed only when they are necessary to ensure driving and traffic-related information does not cause mental overload for people, particularly for onboard people. The key question remaining unsettled is the amount of time for takeover situations. If the time is too short (such as 2 or 3 s) to alert backup drivers or passengers of upcoming danger, it can be stressful for them to understand and react to the situation appropriately. On the other hand, another critical situation may evolve in longer intervals, such as 15-20 s, resulting in late and ineffective human intrusion into that scenario. So, I argue that extensive case-based reasoning and scenario-based evaluations with users in the loop can help determine the best practices with the amount of takeover time. Hence, additional studies should be carried out on the time granularity of AV explanations to reach a consensus on this concept.

9.2.3 Building Robust Explanations

Building robust interactive explanation models for AVs remains yet another challenging task. The experiments described in Chapter 6 and Chapter 8 disclose that even advanced interactive explanation frameworks, particularly transformer-based QA models, are prone to present falsified responses to human adversarial questions. Meanwhile, the user study presented in Chapter 6 and empirical analysis presented

in Chapter 8 reveal that incorrect interactive explanations have a negative impact on users' perceived safety, the feeling of control, and also on the safety of autonomous driving actions, respectively. While transformed-based QA models are increasingly deployed on AVs as conversational models, such issues can significantly damage users' trust and negatively affect regulatory approval of this technology.

As a potential solution, I propose incorporating human-adversarial examples into the training process for driving specific QA models. For instance, when a passenger onboard asks the HMI, "Why is the car turning to the left?" as an adversarial question in an actual *right turn* scenario under a green light, the HMI should present an action-reflecting response like "No, the car is turning to the right as the traffic light allows a right turn," explaining AV action with causal attribution. Regardless of whether such questions are asked intentionally or unintentionally, the explanation models should understand the joint semantics of question and scene and present context-aware responses. Consequently, pre-trained models must be *regulatable* by construction [141] to ensure their inference process leverages human-defined concepts.

9.2.4 Explanations in Uncertainty

Despite substantial advancements in providing a variety of explanations for autonomous driving, conveying the level of confidence to relevant interaction partners in such explanations remains a significant challenge. This issue deepens in uncertainty in the driving environment, such as in adverse weather conditions and situations with reduced visibility (e.g., dense fog, nighttime driving). Without carefully measuring residual risks and environmental uncertainty, overconfident decisions may have dire consequences for an AV and human actors at that traffic scene. Consequently, uncertainty estimation is a vital problem in dealing with unforeseen events safely.

While quantifying uncertainty has recently been investigated by autonomous driving researchers from several aspects, such as for statistical guarantees [185], and object

detection [184, 209], providing relevant explanations with a certain level of confidence under uncertainty relatively remains unexplored in state of the art. Only a few studies have attempted to investigate explainability within uncertainty, such as [203] and [167]; however, the literature is scarce with relevant studies in general. Consequently, as dynamic driving environments often come with considerable uncertainties, there is an imminent need to justify AV actions with a certain level of confidence instead of solely presenting deterministic explanations to people in the loop.

9.2.5 XAI within Level 3 Situation Awareness: Explanations for Projected Events

Currently, AV explanations are primarily leveraged for post-event analysis and real-time action justification and traffic scene understanding, which correspond to Level 1 and Level 2 situation awareness defined by [83], and further proposal from the explainability perspective [233]. It is also interesting to project future AV states and actions based on information currently perceived and comprehended, thereby answering "what-if" questions. Thus, Level 3 explanations in the context of AVs can provide descriptive information about the future projected behavior of an intelligent driving system.

In conclusion, moving forward, I trust that the future of autonomous driving will be shaped by its acceptable safety, transparency, and the rights and responsibilities of its consumers. While these requirements are crucial and a part of today's most recent principles for regulated AI [56], I foresee that the listed directions will not solely remain subjective statements but will be integral components and built-in features of the next-generation, human-friendly AV technology.

Bibliography

[1] R. Abe, "Introducing autonomous buses and taxis: Quantifying the potential benefits in Japanese transportation systems," *Transportation Research Part A: Policy and Practice*, vol. 126, pp. 94–113, 2019.

[2] E. Ackerman, "What Full Autonomy Means for the Waymo Driver," *IEEE Spectrum*, vol. Volume 4, 2021.

[3] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[4] Advanced information and telecommunications network society, *Outline of systematic preparations related to autonomous driving*. The Government of Japan, 2017.

[5] S. Y. Alaba and J. E. Ball, "Transformer-Based Optimized Multimodal Fusion for 3D Object Detection in Autonomous Driving," *IEEE Access*, 2024.

[6] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101 805, 2023.

[7] Y. Almalioglu, M. Turan, N. Trigoni, and A. Markham, "Deep learning-based robust positioning for all-weather autonomous driving," *Nature Machine Intelligence*, vol. 4, no. 9, pp. 749–760, 2022.

[8] M. Althoff and J. M. Dolan, "Online Verification of Automated Road Vehicles Using Reachability Analysis," *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 903–918, 2014.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[10] A. Anderson *et al.*, "Explaining Reinforcement Learning to Mere Mortals: An Empirical Study," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 1328–1334.

[11] Andrew J. Hawkins, "How will driverless cars 'talk' to pedestrians? Waymo has a few ideas," *The Verge*, 2023.

[12] S. Antol *et al.*, "VQA: Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[13] Apex.AI Blog. "An overview of taxonomy, legislation, regulations, and standards for automated mobility." (Accessed on April 8, 2024). (2022), [Online]. Available: https://www.apex.ai/post/legislation-standards-taxonomy-overview.

[14] J. Araluce, L. M. Bergasa, M. Ocaña, Á. Llamazares, and E. López-Guillén, "Leveraging Driver Attention for an End-to-End Explainable Decision-Making From Frontal Images," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[15] S. Arfini, P. Bellani, A. Picardi, M. Yan, F. Fossa, and G. Caruso, "Design for Inclusivity in Driving Automation: Theoretical and Practical Challenges to Human-Machine Interactions and Interface Design," in *Connected and Automated Vehicles: Integrating Engineering and Ethics*, Springer, 2023, pp. 63–85.

[16] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, "Explaining Autonomous Driving Actions with Visual Question Answering," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2023, pp. 1207–1214.

[18] S. Atakishiyev, M. Salameh, and R. Goebel, "Incorporating Explanations into Human-Machine Interfaces for Trust and Situation Awareness in Autonomous Vehicles," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 2948–2955.

[19] S. Atakishiyev, M. Salameh, and R. Goebel, "Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving," *arXiv preprint arXiv:2403.12176*, 2024.

[20] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," *IEEE Access*, vol. 12, pp. 101 603–101 625, 2024.

[21] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards Safe, Explainable, and Regulated Autonomous Driving," in *Explainable Artificial Intelligence for Intelligent Transportation Systems*, CRC Press, 2023, pp. 32–52.

[22] A. Atrey, K. Clary, and D. Jensen, "Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning," in *International Conference on Learning Representations*, 2020.

[23] Austroads, *Guidelines for trials of automated vehicles in Australia*. National Transport Commission, 2020.

[24] E. Awad *et al.*, "The Moral Machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[25] J. Ayoub, N. Du, X. J. Yang, and F. Zhou, "Predicting Driver Takeover Time in Conditionally Automated Driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9580–9589, 2022.

[26] H. Babiker and R. Goebel, "Using KL-divergence to focus Deep Visual Explanation," *31st Neural Information Processing Systems Conference (NIPS), Interpretable ML Symposium. Long Beach, CA, USA*, 2017.

[27] H. K. B. Babiker and R. Goebel, "An Introduction to Deep Visual Explanation," *31st Neural Information Processing Systems Conference (NIPS), Long Beach, CA, USA*, 2017.

[28] C. Badue *et al.*, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113 816, 2021.

[29] P. Bhavsar, P. Das, M. Paugh, K. Dey, and M. Chowdhury, "Risk Analysis of Autonomous Vehicles in Mixed Traffic Streams," *Transportation Research Record*, vol. 2625, no. 1, pp. 51–61, 2017.

[30] I. Bica, D. Jarrett, A. Hüyük, and M. van der Schaar, "Learning "What-if" Explanations for Sequential Decision-Making," in *International Conference on Learning Representations*, 2021.

[31] M. Bojarski *et al.*, "End to End Learning for Self-Driving Cars," *arXiv preprint arXiv:1604.07316*, 2016.

[32] M. Bojarski *et al.*, "VisualBackProp: efficient visualization of CNNs," *arXiv preprint arXiv:1611.05418*, vol. 2, 2016.

[33] C. Brewitt, B. Gyevnar, S. Garcin, and S. V. Albrecht, "GRIT: Fast, Interpretable, and Verifiable Goal Recognition with Learned Decision Trees for Autonomous Driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 1023–1030.

[34] C. Brewitt, M. Tamborski, C. Wang, and S. V. Albrecht, "Verifiable Goal Recognition for Autonomous Driving with Occlusions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 11 210–11 217.

[35] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric Deep Learning: Going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[36] B. Brown and E. Laurier, "The Trouble with Autopilots: Assisted and Autonomous Driving on the Social Road," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 416–429.

[37] Bundesanzeiger Verlag Board, "Act amending the road traffic act and the compulsory insurance act (autonomous driving act)," 2021.

[38] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artificial Intelligence*, vol. 279, p. 103 201, 2020.

[39] Z. Cai, A. Wang, W. Zhang, M. Gruffke, and H. Schweppe, "0-days & Mitigations: Roadways to Exploit and Secure Connected BMW Cars," *Black Hat USA*, vol. 2019, no. 39, p. 6, 2019.

[40] S. Campbell *et al.*, "Sensor Technology in Autonomous Vehicles : A review," in *2018 29th Irish Signals and Systems Conference (ISSC)*, IEEE, 2018, pp. 1–4.

[41] S. Casas, A. Sadat, and R. Urtasun, "MP3: A Unified Model to Map, Perceive, Predict and Plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[42] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable End-to-End Urban Autonomous Driving With Latent Deep Reinforcement Learning," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[43] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end Autonomous Driving: Challenges and Frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[44] L. Chen *et al.*, "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14 093–14 100, 2024.

[45] H.-X. Cheng, X.-F. Han, and G.-Q. Xiao, "TransRVNet: LiDAR Semantic Segmentation With Transformer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 5895–5907, 2023.

[46] F. Chi, Y. Wang, P. Nasiopoulos, and V. C. Leung, "Multi-Modal GPT-4 Aided Action Planning and Reasoning for Self-driving Vehicles," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 7325–7329.

[47] W.-L. Chiang *et al.*, *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*, 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/.

[48] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural Attention Fields for End-to-End Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 793–15 803.

[49] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 878–12 895, 2022.

[50] T. Choudhary *et al.*, "Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16 345–16 352, 2024.

[51] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the Limitations of Behavior Cloning for Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9329–9338.

[52] Comma.AI, *Public driving dataset*, https://github.com/commaai/research, Accessed online on Apr 1, 2024.

[53] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[54] A. Corso and M. J. Kochenderfer, "Interpretable Safety Validation for Autonomous Vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2020, pp. 1–6.

[55] R. C. Coulter, *Implementation of the Pure Pursuit Path Tracking Algorithm*. Carnegie Mellon University, The Robotics Institute, 1992.

[56] Council of the EU, *Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI*, 2024. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/.

[57] C. Cui *et al.*, "A Survey on Multimodal Large Language Models for Autonomous Driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.

[58] J. Cui, L. S. Liew, G. Sabaliauskaite, and F. Zhou, "A review on safety failures, security attacks, and available countermeasures for autonomous vehicles," *Ad Hoc Networks*, vol. 90, p. 101 823, 2019.

[59] Z. Cui, M. Li, Y. Huang, Y. Wang, and H. Chen, "An interpretation framework for autonomous vehicles decision-making via SHAP and RF," in *2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, IEEE, 2022, pp. 1–7.

[60] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining Autonomous Driving by Learning End-to-End Visual Attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 340–341.

[61] Daimler media, *Autonomous concept car smart vision EQ fortwo: Welcome to the future of car sharing*, Accessed on March 10, 2024. [Online]. Available: https://media.mbusa.com/releases/release-80848dccd3f3680a764667ad530987e9-autonomous-concept-car-smart-vision-eq-fortwo.

[62] A. Dandekar, L.-A. Mathis, M. Berger, and B. Pfleging, "How to Display Vehicle Information to Users of Automated Vehicles When Conducting Non-Driving-Related Activities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. MHCI, pp. 1–22, 2022.

[63] David Mullen, *Mercedes to accept legal responsibility for accidents involving self-driving cars*, 2022.

[64] P. De Haan, D. Jayaraman, and S. Levine, "Causal Confusion in Imitation Learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[65] L. De Moura and N. Bjørner, "Z3: An efficient SMT solver," in *Tools and Algorithms for the Construction and Analysis of Systems: 14th International Conference, TACAS 2008*, Springer, 2008, pp. 337–340.

[66] J. DeCastro, K. Leung, N. Aréchiga, and M. Pavone, "Interpretable Policies from Formally-Specified Temporal Properties," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2020, pp. 1–7.

[67] Department for Transport team, *Safe use of automated lane keeping system (alks) summary of responses and next steps*, 2021.

[68] Department of Transportation, *Occupant protection for vehicles with automated driving systems*, https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-03/Final-Rule-Occupant-Protection-Amendment-Automated-Vehicles.pdf, 2022.

[69] H. Detjen, M. Salini, J. Kronenberger, S. Geisler, and S. Schneegass, "Towards Transparent Behavior of Automated Vehicles Design and Evaluation of HUD Concepts to Support System Predictability Through Motion Intent Communication," in *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 2021, pp. 1–12.

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[71] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle," in *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 2021, pp. 1591–1602.

[72] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, p. 101 896, 2023.

[73] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, "Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems," *Transportation Research Part C: Emerging Technologies*, vol. 156, p. 104 358, 2023.

[74] J. Dong, S. Chen, S. Zong, T. Chen, and S. Labi, "Image Transformer for Explainable Autonomous Driving System," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2021, pp. 2732–2737.

[75] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.

[76] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Conference on Robot Learning*, PMLR, 2017, pp. 1–16.

[77] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.

[78] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A Survey of Embodied AI: From Simulators to Research Tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[79] Y. Duan, Q. Zhang, and R. Xu, "Prompting Multi-Modal Tokens to Enhance End-to-End Autonomous Driving Imitation Learning with LLMs," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6798–6805, 2024.

[80] J. Echterhoff, A. Yan, K. Han, A. Abdelraouf, R. Gupta, and J. McAuley, "Driving through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7346–7355.

[81] U. Ehsan and M. O. Riedl, "Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach," in *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, Springer, 2020, pp. 449–466.

[82] U. Ehsan *et al.*, "Operationalizing Human-Centered Perspectives in Explainable AI," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–6.

[83] M. R. Endsley, "Measurement of Situation Awareness in Dynamic Systems," *Human Factors*, vol. 37, no. 1, pp. 65–84, 1995.

[84] M. R. Endsley, "Supporting Human-AI Teams: Transparency, explainability, and situation awareness," *Computers in Human Behavior*, vol. 140, p. 107 574, 2023.

[85] Erez Dagan, *Solving the long-tail with e2e AI: "The revolution will not be supervised"*, 2024. [Online]. Available: https://wayve.ai/thinking/e2e-embodied-ai-solves-the-long-tail/.

[86] S. Faltaous, M. Baumann, S. Schneegass, and L. L. Chuang, "Design Guidelines for Reliability Communication in Autonomous Vehicles," in *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2018, pp. 258–267.

[87] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju, "Examining accident reports involving autonomous vehicles in California," *PLoS One*, vol. 12, no. 9, e0184952, 2017.

[88] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-Language Explanations for Decision Making of Autonomous Driving Based on Semantic Scene Understanding," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[89] Y. Feng and Y. Sun, "PolarPoint-BEV: Bird-eye-view Perception in Polar Points for Explainable End-to-end Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[90] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts?" In *International Conference on Machine Learning*, PMLR, 2020, pp. 3145–3153.

[91] J. Fleetwood, "Public Health, Ethics, and Autonomous Vehicles," *American Journal of Public Health*, vol. 107, no. 4, pp. 532–537, 2017.

[92] P. Foot, *The problem of abortion and the doctrine of double effect*. Oxford, 1967, vol. 5.

[93] B. Friedrich, "The Effect of Autonomous Vehicles on Traffic," *Autonomous driving: Technical, legal and social aspects*, pp. 317–334, 2016.

[94] D. Fu *et al.*, "Drive Like a Human: Rethinking Autonomous Driving with Large Language Models," *arXiv preprint arXiv:2307.07162*, 2023.

[95] GDPR, "Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016," *Official Journal of the European Union*, 2016.

[96] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 3354–3361.

[97] S. Greydanus, A. Koul, J. Dodge, and A. Fern, "Visualizing and Understanding Atari Agents," in *International Conference on Machine Learning*, PMLR, 2018, pp. 1792–1801.

[98] W. Guo, X. Wu, U. Khan, and X. Xing, "EDGE: Explaining Deep Reinforcement Learning Policies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 222–12 236, 2021.

[99] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative Inverse Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[100] L. Hancox-Li, "Robustness in Machine Learning Explanations: Does It Matter?" In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 640–647.

[101] J. P. Hanna *et al.*, "Interpretable Goal Recognition in the Presence of Occluded Factors for Autonomous Vehicles," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 7044–7051.

[102] S. O. Hansson, M.-Å. Belin, and B. Lundgren, "Self-Driving Vehicles—an Ethical Overview," *Philosophy & Technology*, vol. 34, no. 4, pp. 1383–1408, 2021.

[103] S. Hantler and J. King, "An Introduction to Proving the Correctness of Programs," *ACM Computing Surveys*, vol. 8, pp. 331–353, Sep. 1976. DOI: 10.1145/356674.356677.

[104] J. Harris, "The immoral machine," *Cambridge Quarterly of Healthcare Ethics*, vol. 29, no. 1, pp. 71–79, 2020.

[105] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[106] J. Haspiel *et al.*, "Explanations and Expectations: Trust Building in Automated Vehicles," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 119–120.

[107] J. Hawke, V. Badrinarayanan, A. Kendall, *et al.*, "Reimagining an autonomous vehicle," *arXiv preprint arXiv:2108.05805*, 2021.

[108] S. Haykin and R. Lippmann, "Neural networks, A Comprehensive Foundation," *International Journal of Neural Systems*, vol. 5, no. 4, pp. 363–364, 1994.

[109] B. Hejase *et al.*, "Dynamic and interpretable state representation for deep reinforcement learning in automated driving," *IFAC-PapersOnLine*, vol. 55, no. 24, pp. 129–134, 2022.

[110] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 2016, pp. 3–19.

[111] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 241–250.

[112] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowledge-Based Systems*, vol. 214, p. 106 685, 2021.

[113] C. Hewitt, I. Politis, T. Amanatidis, and A. Sarkar, "Assessing public perception of self-driving cars: The autonomous vehicle acceptance model," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 518–527.

[114] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[115] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual Scene Understanding for Autonomous Driving Using Semantic Segmentation," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019, pp. 285–296.

[116] A. Hu *et al.*, "Model-Based Imitation Learning for Urban Driving," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 703–20 716, 2022.

[117] H. Hu, Q. Wang, Z. Zhang, Z. Li, and Z. Gao, "Holistic transformer: A joint neural network for trajectory prediction and decision-making of autonomous vehicles," *Pattern Recognition*, vol. 141, p. 109 592, 2023.

[118] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning," in *European Conference on Computer Vision*, Springer, 2022, pp. 533–549.

[119] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, "Multi-modal Probabilistic Prediction of Interactive Behavior via an Interpretable Model," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 557–563.

[120] Y. Hu *et al.*, "Planning-Oriented Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[121] G. Huang and B. J. Pitts, "Takeover requests for automated driving: The effects of signal direction, lead time, and modality on takeover performance," *Accident Analysis & Prevention*, vol. 165, p. 106 534, 2022.

[122] G. Huang and B. J. Pitts, "The effects of age and physical exercise on multimodal signal responses: Implications for semi-autonomous vehicle takeover requests," *Applied Ergonomics*, vol. 98, p. 103 595, 2022.

[123] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A Survey on Trajectory-Prediction Methods for Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[124] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation Learning: A Survey of Learning Methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[125] IEEE Global Initiative, "A vision for prioritizing human well-being with artificial intelligence and autonomous systems," *IEEE Glob Initiat Ethical Considerations Artif Intell Auton Syst*, vol. 13, 2016.

[126] ISO 21448 Committee. "Road vehicles — Safety of the intended functionality." (Accessed on February 10, 2023). (), [Online]. Available: https://www.iso.org/standard/70939.html.

[127] ISO 26262-6. "ISO 26262-6:2018 Road vehicles — Functional safety — Part 6: Product development at the software level." (Accessed on February 10, 2023). (), [Online]. Available: https://www.iso.org/standard/68388.html.

[128] ISO Technical Commitee, *ISO/SAE 21434:2021 Road vehicles — Cybersecurity engineering*, 2021. [Online]. Available: https://www.iso.org/standard/70918.html.

[129] ISO26262, "ISO26262: Road vehicles-Functional safety," *International Standard ISO/FDIS*, vol. 26262, 2011.

[130] B. W. Israelsen and N. R. Ahmed, ""Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–37, 2019.

[131] M. Itkina and M. Kochenderfer, "Interpretable Self-Aware Neural Networks for Robust Trajectory Prediction," in *Conference on Robot Learning*, PMLR, 2023, pp. 606–617.

[132] ITU Survey Team, *The Molly Problem Public Survey Results (preliminary)*, https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Documents/Survey-Results.pdf?csf=1&e=wb7tAs, 2020.

[133] ITU Team, "The Molly Problem," *AI for autonomous and assisted driving*, (Accessed on February 12, 2024). [Online]. Available: https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx.

[134] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering Counterfactual Explanations with Semantics," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, Springer, 2022, pp. 387–403.

[135] A. Jain, L. Del Pero, H. Grimmett, and P. Ondruska, "Autonomy 2.0: Why is self-driving always 5 years away?" *arXiv preprint arXiv:2107.08142*, 2021.

[136] J. Jarvis Thomson, "The Trolley Problem," *Yale Law Journal*, vol. 94, no. 6, p. 5, 1985.

[137] T. Jing *et al.*, "InAction: Interpretable Action Decision Making for Autonomous Driving," in *European Conference on Computer Vision*, Springer, 2022, pp. 370–387.

[138] T. A. Johansen, T. Perez, and A. Cristofaro, "Ship Collision Avoidance and COLREGS Compliance Using Simulation-Based Control Behavior Selection With Predictive Hazard Assessment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3407–3422, 2016.

[139] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, "Explainable Reinforcement Learning via Reward Decomposition," in *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019.

[140] Keen Security Lab of Tencent, *New Car Hacking Research: 2017, Remote Attack Tesla Motors Again*, 2017.

[141] E. Kenny and J. Shah, "In Pursuit of Regulatable LLMs," in *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[142] E. M. Kenny, M. Tucker, and J. Shah, "Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes," in *The Eleventh International Conference on Learning Representations*, 2023.

[143] T. Kessler *et al.*, "Bridging the Gap between Open Source Software and Vehicle Hardware for Autonomous Driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2019, pp. 1612–1619.

[144] G. Kim, D. Yeo, T. Jo, D. Rus, and S. Kim, "What and When to Explain? On-road Evaluation of Explanations in Highly Automated Vehicles," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–26, 2023.

[145] J. Kim and J. Canny, "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 2961–2969.

[146] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable Learning for Self-driving Vehicles by Internalizing Observation-to-Action Rules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9661–9670.

[147] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual Explanations for Self-Driving Vehicles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 563–578.

[148] J. Kim *et al.*, "Toward explainable and advisable model for self-driving cars," *Applied AI Letters*, e56, 2021.

[149] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Computers & security*, vol. 103, p. 102 150, 2021.

[150] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations*, 2017.

[151] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone, "Reward (mis) design for autonomous driving," *Artificial Intelligence*, vol. 316, p. 103 829, 2023.

[152] S. Kolekar, S. Gite, B. Pradhan, and A. Alamri, "Explainable AI in Scene Understanding for Autonomous Vehicles in Unstructured Traffic Environments on Indian Roads Using the Inception U-Net Model with Grad-CAM Visualization," *Sensors*, vol. 22, no. 24, p. 9677, 2022.

[153] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 9, no. 4, pp. 269–275, 2015.

[154] S. Kothawade, V. Khandelwal, K. Basu, H. Wang, and G. Gupta, "Autodiscern: Autonomous driving using common sense reasoning," *arXiv preprint arXiv:2110.13606*, 2021.

[155] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical Transformer for LiDAR-Based 3D Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 545–17 555.

[156] R. Lanctot *et al.*, "Accelerating the future: The economic impact of the emerging passenger economy," *Strategy analytics*, vol. 5, p. 30, 2017.

[157] P. Langley, "Varieties of explainable agency," in *ICAPS Workshop on Explainable AI Planning*, 2019.

[158] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[159] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[160] D. Lewis, "Causal explanation," 1986.

[161] C. Li, S. H. Chan, and Y.-T. Chen, "Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10 711–10 718.

[162] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[163] T. P. Lillicrap *et al.*, "Continuous Control with Deep Reinforcement Learning," *International Conference on Learning Representations*, 2016.

[164] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 195–204.

[165] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection," *arXiv preprint arXiv:2311.10122*, 2023.

[166] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[167] S. Ling, Y. Wan, X. Jia, and N. Du, "Improving Explainable Object-induced Model through Uncertainty for Automated Vehicles," in *2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 443–451.

[168] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[169] J. Liu, N. Xu, Y. Shi, M. M. Rahman, T. Barnett, and S. Jones, "Do first responders trust connected and automated vehicles (CAVs)? A national survey," *Transport Policy*, vol. 140, pp. 85–99, 2023.

[170] W. Liu, D. Li, E. Aasi, R. Tron, and C. Belta, "Interpretable Generative Adversarial Imitation Learning," *arXiv preprint arXiv:2402.10310*, 2024.

[171] H. Lu, Y. Liu, M. Zhu, C. Lu, H. Yang, and Y. Wang, "Enhancing Interpretability of Autonomous Driving Via Human-Like Cognitive Maps: A Case Study on Lane Change," *IEEE Transactions on Intelligent Vehicles*, 2024.

[172] Q. Lu *et al.*, "KEMP: Keyframe-Based Hierarchical End-to-End Deep Model for Long-Term Trajectory Prediction," in *2022 International Conference on Robotics and Automation (ICRA)*, IEEE, 2022, pp. 646–652.

[173] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[174] B. Lundgren, "Safety requirements vs. crashing ethically: What matters most for policies on autonomous vehicles," *AI & SOCIETY*, vol. 36, no. 2, pp. 405–415, 2021.

[175] A. S. Madhav and A. K. Tyagi, "Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles," in *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, Springer, 2022, pp. 123–136.

[176] S. Magdici and M. Althoff, "Fail-Safe Motion Planning of Autonomous Vehicles," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2016, pp. 452–458.

[177] H. Mankodiya, M. S. Obaidat, R. Gupta, and S. Tanwar, "XAI-AV: Explainable Artificial Intelligence for Trust Management in Autonomous Vehicles," in *2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, IEEE, 2021, pp. 1–5.

[178] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "GPT-Driver: Learning to Drive with GPT," *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[179] A.-M. Marcu *et al.*, "LingoQA: Video Question Answering for Autonomous Driving," *arXiv preprint arXiv:2312.14115*, 2023.

[180] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, "Ethical issues in focus by the autonomous vehicles industry," *Transport reviews*, vol. 41, no. 5, pp. 556–577, 2021.

[181] Mathematics Tutoring at Crafton Hills College, "Standard Normal Distribution Probabilities Table," Accessed on May 15, 2024.

[182] R. McAllister *et al.*, "Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 4745–4753.

[183] A. D. McDonald *et al.*, "Toward Computational Simulations of Behavior During Automated Driving Takeovers: A Review of the Empirical and Modeling Literatures," *Human factors*, vol. 61, no. 4, pp. 642–688, 2019.

[184] G. P. Meyer and N. Thakurdesai, "Learning an Uncertainty-Aware Object Detector for Autonomous Driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10 521–10 527.

[185] R. Michelmore, M. Wicker, L. Laurenti, L. Cardelli, Y. Gal, and M. Kwiatkowska, "Uncertainty Quantification with Statistical Guarantees in End-to-End Autonomous Driving Control," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 7344–7350.

[186] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[187] S. Milani, N. Topin, M. Veloso, and F. Fang, "Explainable Reinforcement Learning: A Survey and Comparative Review," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–36, 2024.

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

[189] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 279–288.

[190] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[191] B. Mok *et al.*, "Emergency, Automation Off: Unstructured Transition Timing for Distracted Drivers of Automated Vehicles," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, IEEE, 2015, pp. 2458–2464.

[192] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2020.

[193] J. M. Müller, "Comparing technology acceptance for autonomous vehicles, battery electric vehicles, and car sharing—A study across Europe, China, and North America," *Sustainability*, vol. 11, no. 16, p. 4333, 2019.

[194] NACTO, *NACTO policy statement on automated vehicles*, https://nacto.org/wp-content/uploads/2016/06/NACTO-Policy-Automated-Vehicles-201606.pdf, Accessed March 10, 2024, 2016.

[195] National Highway Traffic Safety Administration, *Federal automated vehicles policy: Accelerating the next revolution in roadway safety*. US Department of Transportation, 2016.

[196] F. Naujoks, S. Hergeth, K. Wiedemann, N. Schömig, and A. Keinath, "Use cases for assessing, testing, and validating the human machine interface of automated driving systems," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Sage Publications Sage CA: Los Angeles, CA, vol. 62, 2018, pp. 1873–1877.

[197] M. Nie *et al.*, "Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving," *arXiv preprint arXiv:2312.03661*, 2023.

[198] M. L. Olson, R. Khanna, L. Neal, F. Li, and W.-K. Wong, "Counterfactual state explanations for reinforcement learning agents via generative deep learning," *Artificial Intelligence*, vol. 295, p. 103 455, 2021.

[199] D. Omeiza, H. Web, M. Jirotka, and L. Kunze, "Towards Accountability: Providing Intelligible Explanations in Autonomous Driving," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2021, pp. 231–237.

[200] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in Autonomous Driving: A Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 142–10 162, 2021.

[201] U. Onyekpe, Y. Lu, E. Apostolopoulou, V. Palade, E. U. Eyo, and S. Kanarachos, "Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP," in *Explainable AI: Foundations, Methodologies and Applications*, Springer, 2022, pp. 157–183.

[202] R. Paleja, Y. Niu, A. Silva, C. Ritchie, S. Choi, and M. Gombolay, "Learning Interpretable, High-Performing Policies for Autonomous Driving," *Robotics Science and Systems*, 2022.

[203] H. Pan, Z. Wang, W. Zhan, and M. Tomizuka, "Towards Better Performance and More Explainable Uncertainty for 3D Object Detection of Autonomous Vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2020, pp. 1–7.

[204] X. Pan, X. Chen, Q. Cai, J. Canny, and F. Yu, "Semantic Predictive Control for Explainable and Efficient Policy Learning," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 3203–3209.

[205] S. Park *et al.*, "VLAAD: Vision and Language Assistant for Autonomous Driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 980–987.

[206] J. Pearl, *Causality*. Cambridge University Press, 2009.

[207] C. Pek and M. Althoff, "Fail-Safe Motion Planning for Online Verification of Autonomous Vehicles Using Convex Optimization," *IEEE Transactions on Robotics*, vol. 37, no. 3, pp. 798–814, 2020.

[208] C. Pek, S. Manzinger, M. Koschi, and M. Althoff, "Using online verification to prevent autonomous vehicles from causing accidents," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 518–528, 2020.

[209] L. Peng, H. Wang, and J. Li, "Uncertainty Evaluation of Object Detection Algorithms for Autonomous Vehicles," *Automotive Innovation*, vol. 4, no. 3, pp. 241–252, 2021.

[210] X. Peng, M. Riedl, and P. Ammanabrolu, "Inherently Explainable Reinforcement Learning in Natural Language," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 178–16 190, 2022.

[211] Ó. Pérez-Gil *et al.*, "Deep reinforcement learning based control for Autonomous Vehicles in CARLA," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3553–3576, 2022.

[212] J. Petit and S. E. Shladover, "Potential Cyberattacks on Automated Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 546–556, 2014.

[213] D. A. Pomerleau, "ALVINN: An Autonomous Land Vehicle in a Neural Network," *Advances in Neural Information Processing Systems*, vol. 1, 1988.

[214] S. Povolny and S. Trivedi, "Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles," *McAfee Advanced Threat Research*, 2020.

[215] A. Prakash, K. Chitta, and A. Geiger, "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[216] O. Pribyl, R. Blokpoel, and M. Matowicki, "Addressing EU climate targets: Reducing CO2 emissions using cooperative and automated vehicles," *Transportation Research Part D: Transport and Environment*, vol. 86, p. 102 437, 2020.

[217] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing Connected & Autonomous Vehicles: Challenges Posed by Adversarial Machine Learning and the Way Forward," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 998–1026, 2020.

[218] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.

[219] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[220] A. Rasouli and J. K. Tsotsos, "Autonomous Vehicles That Interact With Pedestrians: A Survey of Theory and Practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[221] P. Regulation, "Regulation (EU) 2016/679 of the European Parliament and of the Council," *Regulation (EU)*, vol. 679, 2016.

[222] H. Ren *et al.*, "Graph convolutional networks in language and vision: A survey," *Knowledge-Based Systems*, vol. 251, p. 109 250, 2022.

[223] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "PlanT: Explainable Planning Transformers via Object-Level Representations," in *Conference on Robot Learning*, PMLR, 2023, pp. 459–470.

[224] A. Rezaei and B. Caulfield, "Safety of autonomous vehicles: What are the insights from experienced industry professionals?" *Transportation research part F: traffic psychology and behaviour*, vol. 81, pp. 472–489, 2021.

[225] G. Rjoub, J. Bentahar, and O. A. Wahab, "Explainable AI-based Federated Deep Reinforcement Learning for Trusted Autonomous Driving," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, IEEE, 2022, pp. 318–323.

[226] P. E. Ross, "The Audi A8: the World's First Production Car to Achieve Level 3 Autonomy," *IEEE Spectrum*, vol. 1, 2017.

[227] T. R. Roth-Berghofer, "Explanations and case-based reasoning: Foundational issues," in *European Conference on Case-Based Reasoning*, Springer, 2004, pp. 389–403.

[228] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[229] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[230] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110 273, 2023.

[231] R. Salay, R. Queiroz, and K. Czarnecki, "An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software," *Safety of the Intended Functionality*, pp. 13–25, 2020.

[232] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, "Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation," *arXiv preprint arXiv:1611.08191*, 2016.

[233] L. Sanneman and J. A. Shah, "The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems," *International Journal of Human-Computer Interaction*, vol. 38, no. 18-20, pp. 1772–1788, 2022.

[234] A. Schieben, M. Wilbrink, C. Kettwich, R. Madigan, T. Louw, and N. Merat, "Designing the interaction of automated vehicles with other traffic participants: Design considerations based on human needs and expectations," *Cognition, Technology & Work*, vol. 21, pp. 69–85, 2019.

[235] L. M. Schmidt, G. Kontes, A. Plinge, and C. Mutschler, "Can You Trust Your Autonomous Car? Interpretable and Verifiably Safe Reinforcement Learning," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2021, pp. 171–178.

[236] T. Schneider, S. Ghellal, S. Love, and A. R. Gerlicher, "Increasing the User Experience in Autonomous Driving through different Feedback Modalities," in *26th International Conference on Intelligent User Interfaces*, 2021, pp. 7–10.

[237] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, and A. R. Gerlicher, "ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12.

[238] T. Schneider *et al.*, "Don't fail me! The Level 5 Autonomous Driving Information Dilemma regarding Transparency and User Experience," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 540–552.

[239] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[240] S. K. Seyed Ghasemipour, S. S. Gu, and R. Zemel, "SMILe: Scalable Meta Inverse Reinforcement Learning through Context-Conditional Policies," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[241] H. Sha *et al.*, "LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving," *arXiv preprint arXiv:2310.03026*, 2023.

[242] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-Enhanced Autonomous Driving Using Interpretable Sensor Fusion Transformer," in *Conference on Robot Learning*, PMLR, 2023, pp. 726–737.

[243] Y. Shen, S. Jiang, Y. Chen, and K. R. Driggs-Campbell, "To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles," in *NeurIPS Workshop on Progress and Challenges in Building Trustworthy Embodied AI*, 2022.

[244] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," in *International Conference on Machine Learning*, PMLR, 2017, pp. 3145–3153.

[245] J. Shuttleworth, *SAE standard news: J3016 automated-driving graphic update*, https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic, 2019.

[246] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2015.

[247] S. Sohrabi, H. Khreis, and D. Lord, "Impacts of autonomous vehicles on public health: A conceptual model and policy recommendations," *Sustainable cities and society*, vol. 63, p. 102 457, 2020.

[248] Z. Song *et al.*, "An Interpretable Deep Reinforcement Learning Approach to Autonomous Driving," in *IJCAI Workshop on Artificial Intelligence for Autonomous Driving*, 2022.

[249] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[250] N. A. Stanton, P. M. Salmon, G. H. Walker, and M. Stanton, "Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian," *Safety Science*, vol. 120, pp. 117–128, 2019.

[251] J. Suchan, M. Bhatt, and S. Varadarajan, "Out of sight but not out of mind: An answer set programming based online abduction framework for visual sense-making in autonomous driving," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 1879–1885.

[252] T. Sun *et al.*, "SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 371–21 382.

[253] X. Sun, F. R. Yu, and P. Zhang, "A Survey on Cyber-Security of Connected and Autonomous Vehicles (CAVs)," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6240–6259, 2021.

[254] Z. N. Sunberg, C. J. Ho, and M. J. Kochenderfer, "The value of inferring the internal state of traffic participants for autonomous freeway driving," in *2017 American control conference (ACC)*, IEEE, 2017, pp. 3004–3010.

[255] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second Edition. MIT Press, 2018.

[256] H. Tabani, L. Kosmidis, J. Abella, F. J. Cazorla, and G. Bernat, "Assessing the Adherence of an Industrial Autonomous Driving Framework to ISO 26262 Software Guidelines," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[257] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A Survey of End-to-End Driving: Architectures and Training Methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1364–1384, 2020.

[258] S. Tekkesinoglu, A. Habibovic, and L. Kunze, "Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap," *arXiv preprint arXiv:2404.00019*, 2024.

[259] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical Interpretable Imitation Learning for End-to-End Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 673–683, 2022.

[260] The High-Level Expert Group on AI at the European Commission, *Ethics guidelines for trustworthy AI*, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, Accessed on March 11, 2024, 2019.

[261] S. Thrun *et al.*, "Stanley: The robot that won the DARPA Grand Challenge," *Journal of Field Robotics*, vol. 23, no. 9, pp. 661–692, 2006.

[262] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Cooperative Autonomous Vehicles that Sympathize with Human Drivers," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2021, pp. 4517–4524.

[263] Transport Canada, "Guidelines for testing automated driving systems in Canada," *The Ministry of Transportation of Canada*, 2021.

[264] Udacity, *Public driving dataset*, https://public.roboflow.com/object-detection/self-driving-car, Accessed online on Apr 6, 2024.

[265] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[266] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[267] S. Vom Dorff, B. Böddeker, M. Kneissl, and M. Fränzle, "A fail-safe architecture for automated driving," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2020, pp. 828–833.

[268] G. A. Vouros, "Explainable Deep Reinforcement Learning: State of the Art and Challenges," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–39, 2022.

[269] J. Wan and C. Wu, "The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 6, pp. 582–591, 2018.

[270] C. Wang, T. H. Weisswange, M. Krueger, and C. B. Wiebel-Herboth, "Human-Vehicle Cooperation on Prediction-Level: Enhancing Automated Driving with Human Foresight," in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, IEEE, 2021, pp. 25–30.

[271] C. Wang and N. Aouf, "Explainable Deep Adversarial Reinforcement Learning Approach for Robust Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, 2024.

[272] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing Theory-Driven User-Centric Explainable AI," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[273] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning Interpretable End-to-End Vision-Based Motion Planning for Autonomous Driving with Optical Flow Distillation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 13 731–13 737.

[274] H. Wang, W. Wang, S. Yuan, and X. Li, "Uncovering Interpretable Internal States of Merging Tasks at Highway on-Ramps for Autonomous Driving Decision-Making," *IEEE Transactions on Automation Science and Engineering*, 2021.

[275] L. Wang, X. Zhang, H. Su, and J. Zhu, "A Comprehensive Survey of Continual Learning: Theory, Method and Application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[276] Waymo Team, *Self-Driving Car Technology for a Reliable Ride*, 2024.

[277] Wayve Blog, *LINGO-1: Exploring Natural Language for Autonomous Driving*, https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/, 2023.

[278] Wayve Blog, *LINGO-2: Driving with Natural Language*, 2024.

[279] L. Wen *et al.*, "DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models," *International Conference on Learning Representations*, 2024.

[280] M. R. Wicker, J. Heo, L. Costabello, and A. Weller, "Robust Explanation Constraints for Neural Networks," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[281] G. Wiegand, M. Eiband, M. Haubelt, and H. Hussmann, ""I'd like an Explanation for That!" Exploring Reactions to Unexpected Autonomous Driving," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–11.

[282] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, and H. Hussmann, "I drive-you trust: Explaining driving behavior of autonomous cars," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[283] W. Woods, J. Chen, and C. Teuscher, "Adversarial explanations for understanding image classification decisions and improved neural network robustness," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 508–516, 2019.

[284] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting Driver Attention in Critical Situations," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, Springer, 2019, pp. 658–674.

[285] Y. Xu *et al.*, "Explainable Object-Induced Action Decision for Autonomous Vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9523–9532.

[286] Z. Xu *et al.*, "DriveGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model," *IEEE Robotics and Automation Letters*, 2024.

[287] W. Xue, Z. Wang, R. Zheng, X. Mei, B. Yang, and K. Nakano, "Fail-Safe Behavior and Motion Planning Incorporating Shared Control for Potential Driver Intervention," *IEEE Transactions on Intelligent Vehicles*, 2023.

[288] Q. Yang, H. Wang, M. Tong, W. Shi, G. Huang, and S. Song, "Leveraging Reward Consistency for Interpretable Feature Discovery in Reinforcement Learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023.

[289] H. Yao, C. Szepesvári, B. A. Pires, and X. Zhang, "Pseudo-MDPs and Factored Linear Action Models," in *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, IEEE, 2014, pp. 1–9.

[290] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[291] F. Yu *et al.*, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2636–2645.

[292] J. Yuan *et al.*, "RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model," *Robotics: Science and Systems*, 2024.

[293] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-Channel Transformer for 3D Lidar-Based Video Object Detection for Autonomous Driving," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2068–2078, 2021.

[294] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[295] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *International Journal of Computer Vision*, vol. 130, no. 10, pp. 2425–2452, 2022.

[296] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.

[297] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "OCTET: Object-aware Counterfactual Explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 062–15 071.

[298] W. Zeng *et al.*, "End-to-End Interpretable Neural Motion Planner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8660–8669.

[299] C. Zhang *et al.*, "Rethinking Closed-Loop Training for Autonomous Driving," in *European Conference on Computer Vision*, Springer, 2022, pp. 264–282.

[300] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via Decision Trees," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6261–6270.

[301] Z. Zhang, R. Tian, R. Sherony, J. Domeyer, and Z. Ding, "Attention-Based Interrelation Modeling for Explainable Automated Driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[302] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[303] X. Zhou *et al.*, "Vision Language Models in Autonomous Driving: A Survey and Outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.