Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions

Shahin Atakishiyev^a, Mohammad Salameh^b, Hengshuai Yao^{a,c}, Randy Goebel^a

^aDepartment of Computing Science, University of Alberta, Edmonton, AB, Canada ^bHuawei Technologies Canada, Co., Ltd., Edmonton, AB, Canada ^cSony AI, Edmonton, AB, Canada

Abstract

Autonomous driving has achieved significant milestones in research and development over the last decade. There is increasing interest in the field as the deployment of self-operating vehicles promises safer and more ecologically friendly transportation systems. With the rise of computationally powerful artificial intelligence (AI) techniques, autonomous vehicles can sense their environment with high precision, make safe real-time decisions, and operate reliably without human intervention. However, intelligent decisionmaking in autonomous cars is not generally understandable by humans in the current state of the art, and such deficiency hinders this technology from being socially acceptable. Hence, aside from making safe real-time decisions, the AI systems of autonomous vehicles also need to explain how their decisions are constructed in order to be regulatory compliant across many jurisdictions. Our study sheds comprehensive light on the development of explainable artificial intelligence (XAI) approaches for autonomous vehicles. In particular, we make the following contributions. First, we provide a thorough overview of the stateof-the-art studies on XAI for autonomous driving. We then propose an XAI framework that considers the societal and legal requirements for the explainability of autonomous driving systems. Finally, as future research directions, we provide several XAI approaches that can improve operational safety and transparency to support public approval of autonomous driving technology by regulators and engaged stakeholders.

Keywords: Explainable artificial intelligence, autonomous driving, intelligent transportation systems, regulatory compliance

1. Introduction

A survey of the American National Highway Traffic Safety Administration (NHTSA) reports that nearly 94% of road accidents are due to human errors [1]. These human-related mistakes are mainly classified as driver distraction, drunk or otherwise impaired driving, lack of attention, violation of the traffic rules, limited view of traffic conditions, and jay-walking pedestrians [2]. The lack of rule obedience, the increasing number of vehicles on roads, and improper road culture have therefore motivated officials, manufacturers, and legislators to make substantial improvements in transportation systems. There are growing research and development attempts to enhance safety and automation capability of autonomous vehicles (AVs), prevent traffic accidents, and create a better road infrastructure. The potential benefits of AVs are improved convenience, operational safety (especially for seniors and people with reduced mobility) [3], reduced CO_2 emissions [4], diminished transportation costs [5], improved safety [6, 7], and reduced traffic density [8]. In particular, reduced traffic congestion and safety assurance are two significant promises of autonomous vehicles. Intel's report on the projected benefits of self-driving cars



Figure 1: A canonical exemplar of explainable AI in autonomous driving: A vehicle provides a natural and intelligible explanation of its real-time decision to bystanders. Graphics adapted and modified from the source [9].

estimates that deployment of this technology 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 [10]. While the potential impact and benefits of automated vehicles in everyday life are promising, there is a major societal concern about the reliability 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 inappropriate autonomous decisions [11, 12, 13, 14]. As AI approaches provide the foundation for real-time driving actions and operations, engaged consumers and regulatory organizations analyze the intelligent driving system of a vehicle to comprehend whether inappropriate decisions of a car are the actual cause of accidents. Therefore, there is an inherent need and expectation from consumers and regulators that AI-driven operations of AVs should be explainable ¹ (e.g., Figure 1) to confirm operational safety. In a recent study, the authors have proposed a framework that describes the fundamental concepts and process steps associated with XAI-based autonomous driving [15]. In this study, we extend the scope of the mentioned work by discussing the following research questions:

- 1. Why is there a need for XAI in autonomous driving technology?
- 2. How do industrial priorities inform the choice of research directions?
- 3. What are the current regulatory requirements directing research priorities in the co-development of autonomous driving architectures and their explanatory components?

With these focus points, our paper makes the following contributions:

- We provide a survey of state-of-the-art XAI-based studies for autonomous driving.
- We present a general design framework for explainable autonomous driving.
- We propose future research directions on XAI approaches for autonomous driving with the goal of ensuring public trust and approval.

¹We'll use the terms explainable and interpretable interchangeably.

The rest of the article is structured as follows. In Section 2, we provide background information and the factors triggering the need for the emergence of XAI in autonomous driving. We then describe the concept of explanations in autonomous driving by why they are needed, to whom they are addressed, and how to construct explanations in Section 3. We present a concise overview of common AI approaches powering the real-time decisive actions of autonomous vehicles in Section 4. In Section 5, we provide a comprehensive survey of state-of-the-art investigations on explainable AI-based autonomous driving. Motivated by current limitations and trends delineated in these works, we present a general explainable autonomous driving framework that considers safety, public expectations, and regulatory principles for the design and development of a learning software architecture of intelligent driving systems. Finally, we provide a future perspective of XAI approaches in autonomous driving and sum up the article.

2. Background

2.1. AVs operations

Autonomous vehicles are systems capable of sensing their environment and mapping such sensing data to real-time driving decisions using an intelligent driving system. To discern, identify, and distinguish the objects in their operational surroundings, autonomous vehicles fuse information from a variety of sensors that help make real-time driving decisions [16, 17]. Current autonomous vehicles deployed on road networks have different levels of automation based on their in-vehicle technologies and intelligent capabilities. SAE International (previously known as the Society of Automotive Engineers) has defined six levels of autonomous driving [18]: Level 0 - No automation (a human driver is responsible 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 in this level); Level 3 - Conditional automation (a vehicle has more advanced features such as object/obstacle detection and can carry out the most driving operation); 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). Real-time decision-making for autonomous vehicles involves several interconnected operational stages. These operations are commonly categorized as perception, localization, planning, and control (Figure 2). Perception is the sensing of an operational environment defined as a combination of two tasks: road surface extraction and on-road object detection [19]. Information for the purpose of perception can be obtained from multi-modal data sources, which currently include LIDAR, RADAR, visible spectrum cameras, and ultrasonic sensors [17, 20]. The process of localization enables an autonomous vehicle to accurately determine its position in the sensor model of the world [21, 22]. The most effective way to get a position of autonomous vehicles is to use satellite navigation-based systems. Among such systems, the Global Navigation Satellite System (GNSS) and its most popular instance, Global Positioning System (GPS), is a universal sensor to determine a global location of a car [23]. As the autonomous car perceives its surroundings and gets its precise localization, it plans the trajectory from the initial point to the final destination. Planning is a complex operation that integrates components of route planning (i.e., selection of a route in the road work from the initial point to the final destination) and behavior planning (such as interactions with other vehicles, people that may be met on a trajectory). Finally, control of an autonomous vehicle is the appropriate execution of planned motions. Feedback controllers mainly manage this function; in modern autonomous vehicles, control is typically carried out through the ADAS software. These systems interact with the sensors of an environment and assist the car in controlling its trajectory along the journey [24]. The currently deployed examples of ADAS include adaptive cruise control, anti-lock braking system, collision avoidance systems, forward collision warning, and lane departure warning systems [25]. So, autonomous vehicles can operate on roads without

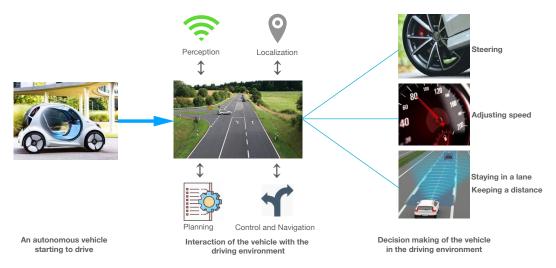


Figure 2: A general structure of AI-driven decision making of an autonomous vehicle in its driving environment.

human assistance by fusing information from multi-modal data sources with AI-powered computer vision, decision-making, and control algorithms.

2.2. Existing issues

Artificial intelligence 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. As the AI-powered driving systems of vehicles advance, the number of autonomous vehicles deployed to road networks has proliferated significantly in many developed European countries, the US, and Canada over the last decade [26]. 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 autonomous cars on roads. For example, Fleetwood [27] 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 [28, 29] (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.). Another critical aspect of the work is a concern for the potential rights and liabilities of passengers sitting inside an autonomous car (for example, by using such a car, does a passenger agree to face potential risks; does a passenger have the responsibility and liability to protect other road users if an accident happens?). That study concludes with four directions - clear and cross-disciplinary discussions amongst stakeholders, including a driving system's action planning choice of an algorithm; enhancing society's knowledge on the issues and limitations of autonomous driving; confirming society's knowledge on solutions of the current issues and proper use of autonomous vehicles; and developing faithful, rational, and monitored standards for public health experts' attention.

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 [30], 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 [31] which force deeper consideration of such dilemmas [32]. Sohrabi et al. [33] have explored the outcomes

of autonomous driving technology on public health in an urban area. They determined that this technology can affect public health in thirty-two directions: more than half of these pathways (seventeen) are negative, eight are positive, and the remaining ones are unsettled. Another similar study led by Martinho et al. [34] has inspected autonomous vehicle-related ethics in both scientific literature and various industry reports in California. According to this survey, both the scholarly literature and industry documents highlight safety and cybersecurity as main issues and underline concerns for moral decision-making algorithms, human-involved control, ethics, design, privacy, accountability, and sustainability, among other concerns. Burton et al. [35] 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 is 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*. That study also shows that the core of these issues is associated with domain complexity, system complexity, and transferring more decision-making functions from humans to autonomous systems. In a relevant case study, authors relate unpredictable urban environments, driving system complexity, and lack of human override in highly autonomous driving to the enumerated core reasons, respectively. Finally, several studies have attempted to elucidate other related issues such as privacy [36, 37], design and implementation issues [38], legal regulation [39], user concerns [40], IoT challenges [41], integration in smart cities [42], as well as management [43] and security concerns [44]. The key findings outlined in the above studies require an understanding of the causes of these issues and intrinsically give the stakeholders the right to ask "why" questions. So, we immediately observe an immense need for explanations about the performance of self-driving cars. Providing explanations of critical decisions can significantly increase the acceptance of autonomous vehicles by both the transportation jurisdictions and community. In the subsequent sections, we discuss established standards and regulations for autonomous driving technology and the appropriate development of AI software architecture for it. We then provide the need for explanations in AVs, which necessitates the emergence of explainable AI methods for intelligent vehicles.

2.3. AVs 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 the development of 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 [45]. 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 [46]. Similarly, Mohseni et al. [47] have tabled a broad description of ethical AI, with a consequent impact on autonomous systems. In this context, autonomous vehicle systems also need to comply with these rules, principles, and requirements. As per the guidelines, the intelligent driving system of the autonomous vehicle should be able to provide intelligible explanations to the engaged stakeholders on the decisions and actions of a car in confirming the safety of autonomous systems and in support of an investigation of road accidents and other critical conditions.

Various organizations have recently proposed guidelines on the regulation of autonomous vehicles to monitor their compliance with law enforcement. NACTO's (National Association of City Transportation Officials) statement on automated vehicles [48] proposes nine principles to shape a policy on regulation of future generation autonomous vehicles. Another well-defined guideline, The Research and Development

(RAND) Corporation's principles, covers promises and issues of autonomous vehicles, and an association of this technology to law and liability issues. Their principles also provide thorough guidance for public regulators to investigate transportation accidents and make safety recommendations (e.g., [49, 50]) to the regulators and manufacturers of autonomous vehicles such as the American National Highway Traffic Safety Administration (NHTSA), SAE International, Tesla, and Apple. NHTSA of the US Department of Transportation has a specific federal guideline on automated vehicle policy to improve traffic safety [51] In March 2022, NHTSA announced that automobile manufacturers would no longer have to equip fully autonomous cars with manual control elements, such as steering wheel and braking pedals in the USA [52]. The Government of Canada has also recently released their comprehensive federal guidelines on testing and regulations of automated driving systems [53]. Their documentation provides detailed information and a regulatory road map for the relevant organizations on the engagement with government agencies, pre-trial, testing, and post-test considerations of autonomous vehicles. In another recently adopted regulation, Germany has published an act on operations of driverless cars, particularly relevant to designated areas of public roads [54]. The UK government has also advanced their interests toward regulated and safe autonomous driving, and hands-free driving was expected to be legally allowed there by the end of 2021 [55]. Other developed countries such as Australia [56], and Japan [57] have also recently launched initiatives for trials of 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 [58], which specifies situational awareness standards to maintain operational safety under the "Safety of the Intended Functionality," and the ISO 26262 [59] standard defined for the safety of electrical and electronic systems in production passenger vehicles, entitled as "Road vehicles - Functional safety." In this context, ISO/TC 204 [60] is the primary standard that provides a comprehensive guide on the overall system and infrastructure aspects of intelligent transportation systems (ITS), supporting the standardization of autonomous driving technology. Motivated by ISO/TC 204, some regional initiatives have also imposed relevant standards on the regulation of autonomous vehicles. For instance, The European Committee for Standardization, together with ISO, has the CEN/TC 278 [61] standard that develops acceptable levels of quality, use cases, and best practices for ITS in Europe. It turns out that autonomous vehicles, or ITS, as a more general field, involve many multidisciplinary foundations to meet the involved stakeholders, insurance, and law enforcement requirements. Thorough documentation on the details of legislation, regulation, and standardization of automated vehicles can be viewed here [62].

3. Explanations in autonomous driving

3.1. Why?

As can be inferred from the above discussion, the need for explanations in autonomous driving arises from existing issues, established regulations and standards covered in previous subsections, and from cross-disciplinary views and opinions of society. At the highest level, the need for explanations in autonomous driving systems can be summarized in terms of three - *psychological*, *sociotechnical*, and *philosophical* perspectives. While traffic accidents and safety concerns remain the main cause of the need for XAI in autonomous driving from a psychological view, from the sociotechnical lens, the key idea is that the design, development, and deployment of autonomous vehicles should be human-centered. As humans are the main social actors and users of this technology, the development principles of AVs should reflect the target audience's needs and take their prior opinions and expectations into account [63, 64]. From a philosophical point of view, explaining AI decisions can provide descriptive information about the causal history of actions taken, particularly in critical situations [65, 66, 67]. Considering these multi-dimensional perspectives, explainable autonomous driving can bring the following benefits to the stakeholders:

• Human-Centered Design

Getting the end users' inputs, opinions, and anticipations on the design and development of the semi or fully AVs will increase the acceptance of this technology by the community [68]. In this context, it is essential to provide user-friendly human-computer interaction and interfaces to the stakeholders, such as backup drivers and passengers. As a solid example, a self-driving vehicle may provide a user interface for in-vehicle passengers or backup drivers on decisive actions. There have also been several studies that use human-centered XAI design in the forms of light, visual, audio, and textual information to transmit the instant decisions of a car to the in-vehicle passengers and drivers [69, 70, 71]. Schneider et al.'s recent empirical study confirms that applying stakeholders' multi-modal feedback to the simulated design of autonomous driving creates a positive user experience [72]. So, a human-centered design that uses intelligible AI methods is a necessary step for the widespread adoption of AV technology.

Trustworthiness

As careless and hazardous driving can directly impact the safety of passengers and bystanders, people naturally require confirmation of the safety of transportation systems. In addition, understanding the causes of actions or decisions is a natural human requirement. As stated in Riberio et al.'s [73] work, "if the users do not trust a model or a prediction, they will not use it." In their case study, Holliday et al. [74] have also empirically shown that providing explanations and perceptible systems significantly increases users' trust in a system. On the other hand, frequently occurring failures without faithful explanations to stakeholders can seriously damage individual and public trust in intelligent systems. Once trust in an intelligent system is damaged, regaining it can be onerous [75]. Israelson and Ahmed [76] have shown that there is an inherent need for algorithmic assurance to build trust in human-autonomous system relationships in their detailed analysis. Therefore, constructing explainable intelligent driving systems is a viable promise for the trustworthy use of AVs.

· Transparency and accountability

If trustworthiness is developed for the intelligent decision-making of a car, it further brings transparency and accountability to AVs technology. Martinho et al. [34] have noted that accountability combines liability and responsibility as a broader concept. In the context of autonomous driving, accountability can be defined as compliance with established legal principles in specific jurisdictions. As the regulatory standards urge the "right to an explanation" as required by GDPR [45], accountability becomes a crucial concept that combines social expectations and legislative norms on the autonomous driving spectrum. In addition, achieving accountability also helps to deal with potential liability and responsibility gaps, as defined by Burton et al., [35], in potential post-accident investigations with the involvement of autonomous cars. Recently Mercedes-Benz has 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 [77]. Mercedes's declaration of legal culpability is a significant milestone toward the accountability of AVs technology.

3.2. *To whom?*

The details, types, and delivery of explanations vary in accordance with users' identities and background knowledge in autonomous driving. For instance, a user having little technical expertise on how autonomous vehicles operate may be satisfied with a simple explanation of a relevant decision/outcome. However, an autonomous systems engineer will need more informative explanations to understand the current operability of the car, with the motivation to appropriately "debug" the existing system as required. Therefore, the use of domain knowledge and expertise of the explainee is essential to provide pertinent,

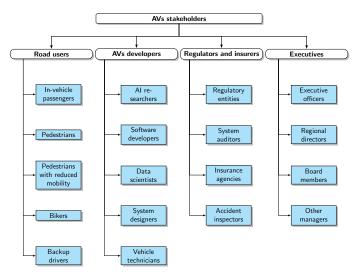


Figure 3: Taxonomy of the stakeholders in autonomous driving.

sufficiently informative, and intelligible explanations [78, 79]. Motivated by a target audience definition of [80] and [81], we can distinguish *four* groups of the stakeholders in autonomous driving, namely Group 1 - Road users, Group 2 - AVs developers, Group 3 - Regulators and insurers, and Group 4 - Executive management of automobile companies. Figure 3 provides the identity of such stakeholders and their positions in the corresponding classification.

3.3. How?

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 explainable AI research. Zablocki et al. [82] define four "W" questions in explainable AI-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 applied explanations to automated collaborative filtering systems [83] and knowledge-intensive case-based reasoning systems [84]. Another empirical approach attempted to derive explanations based on some intelligibility types [85] and used "why," "why not," "what if," and "how to" type explanations for causality filtering. In a recent study, Liao et al. [86] 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.

Another popular approach to produce explanations is based on using psychological tools from *for-mal theories*, according to the literature review of [87]. 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. Like their derivation type, explanations also differ depending on the class in which they are included. Through their extensive survey, Omeiza et al. [81] propose the following dimensions of explanations in the context of autonomous driving:

Explanations based on cause filters: Based on available knowledge, explanations use predefined causes to explain the outcome of an event. The explanations are generated based on cause filters such as "why," "why not," "how to," and "what if" queries (e.g., "Why did the car take the left lane instead of the right lane?"). In fact, we can note that this kind of explanations can be used across many autonomous driving operations.

Explanations based on content type: In this category, explanations are classified based on the components or elements involved with the explanations and the way they are presented. Examples of content types include input influence, input sensitivity, case bases, and demographic factors (for instance, explanations based on what input variables (i.e., driving features) contribute more to the predictive actions).

Explanations depending on a model: Here, explanations are distinguished by being either *model agnostic* or *model dependant*. For instance, some autonomous driving operations can be condition-specific and some can be general, regardless of the driving conditions (for example, explaining *any* autonomous driving action, as a model-agnostic rationale, can be specified for this group of explanations).

Explanations based on a system type: This category attempts to capture the properties of the operational system: [81] distinguish two kinds of explanations as either *data-driven* (i.e., explaining the outcome of a predictive model) or *goal-driven* (explaining an agent's behaviors based on achieving its goal in a predefined setting).

Explanations with interactivity: Once an explanation is provided, a user may further ask follow-up questions to further understand a provided explanation. This feature brings interactivity into the explanation framework (e.g., a user interface for in-vehicle passengers that provides real-time explanations for corresponding actions).

Explanations with concrete scope: This category captures the feasibility and range of explanations that the system can generate by being either *local* or *global*. Local explanations are limited to explanations on some or a subset of all possible actions (i.e., explaining a single prediction in a specific traffic scenario). Global explanations, on the other hand, are capable of explaining all high-level decisions from an initial point to the destination, such as why an autonomous car chose a specific map, why it changed the planned route in the middle of the travel, and so on.

4. AI for autonomous driving

Real-time decisions in autonomous driving are based on environmental perception, processing temporal sequence data, and mapping of real-time perception to relevant actions. In this regard, before reviewing relevant XAI techniques, we provide a concise overview of three major AI approaches to the development of autonomous driving control: convolutional neural networks, recurrent neural networks, and reinforcement learning. These three broad categories of methods focus on mapping sensory information to appropriate actions. The stated AI architectures and their enhancements are dominant techniques that have proven empirical successes in the development and deployment of state-of-the-art AVs.

4.1. Convolutional neural networks

Convolutional neural networks (CNN) are an AI architecture typically used to process spatial information, such as images and videos [88]. 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 approximators. The input x of each layer in a CNN model is organized in 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:

$$\theta = [W, b] \tag{1}$$

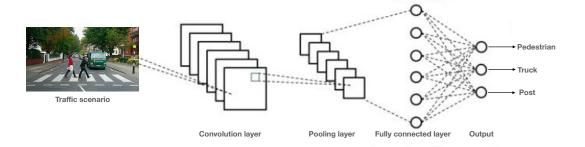


Figure 4: An example of a CNN for object classification in a real-time traffic scenario.

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. Commonly used activation functions with CNN are Sigmoid, Tanh, and ReLu functions while ReLu is the most preferred, because of its relatively faster convergence. The network then makes a final prediction. 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 4.

4.2. Recurrent neural networks

Recurrent neural networks (RNNs) are a deep learning architecture designed for processing temporal and sequential data, such as time series, video, and natural language data [89, 90]. 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

$$h_t = f_H(o_t), (2)$$

in which

$$o_t = W_{IH}x_t + W_{HH}h_{t-1} + b_h. (3)$$

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:

$$y_t = f_O(W_{HO}h_t + b_o). (4)$$

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. One major problem with traditional RNNs is the so-called vanishing gradient during training: network weights might not be effectively updated if the network is very deep. This may result in

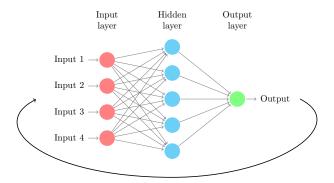


Figure 5: A structure of a typical recurrent neural network

considerably small weight values that may reduce the network's learning ability. To solve this problem, long short-term memory (LSTM), as an enhancement of RNN, can handle sequential data more effectively and learn better [91]. Compared to a simple RNN architecture, LSTM has "gates" that control the flow of information through the network. With this augmented learning capability, LSTMs are more practical than traditional RNNs. With the inherent ability of learning sequential data, RNN and its augmented forms, such as LSTM, can be used to predict future position, velocity, and other parameters of autonomous driving. Figure 5 shows a structure of a typical recurrent neural network.

4.3. 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 by such an interaction [92]. 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 reinforcement learning (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:

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

$$:= V_{\pi}(s)$$
(5)

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:

$$Q_{\pi}(s,a) = \mathbb{E}_{\pi} \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} \mid s_0 = s, a_0 = a \right\}.$$
 (6)

Basically, the value function is a measure of how good it is for an agent to be in a particular state [92]. 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 and observe all the states of the operational environment. In such a situation, an agent's interaction with the surrounding is constructed as a partially observable Markov decision process (POMDP). In the POMDP setting, states are replaced by *observations* [93]. 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* [92]. 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 an autonomous vehicle's real-time decisions. A general diagram of RL for autonomous driving is described in Figure 6.



Figure 6: A diagram of reinforcement learning for a self-driving vehicle.

5. XAI for autonomous driving: A survey on the state of the art

Motivated by the current limitations of AVs technology from an explainability perspective, there have been substantial efforts to build intelligent driving systems that generate intelligible explanations on a vehicle's decisive actions. In general, three types of explanations are common for self-driving vehicles' actions: *visual*, *textual* justifications, and *feature importance* scores. At the highest level, visual explanations are about understanding which portions of an image influence a vehicle controller to take particular actions. Textual explanations aim to provide intelligible rationales behind the actions taken by a vehicle, using a natural and understandable language. Finally, feature importance scores indicate how much each input feature contributes to the prediction of the model quantitatively. We review these investigations that generate some form of explanation for autonomous driving tasks and systematically overview them with their delivery types.

5.1. Vision-based 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 achieve visual explanations. In this regard, explainable CNN architectures have resulted in adjustments to generate visual explanations. Zeiler et al. [94] use deconvolution layers to understand the internal representation of CNNs in their seminal work. Hendricks et al. [95] propose a model concentrating on distinguished properties of objects that explain the rationale for the predicted label. Zhou et al.'s [96] saliency map architecture, class activation map (CAM), highlights the discriminative

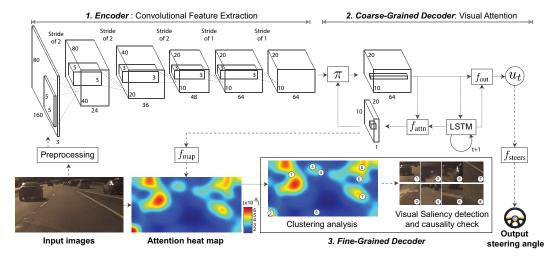


Figure 7: End-to-end learning of steering angle commands from an input image. Source: [98].

part of an image to predict the label of the image. Moreover, Selvaraju et al. [97] 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, [99], layerwise relevance propagation [100, 101], and DeepLift [102]. Babiker and Goebel [103, 104] have shown that heuristics-based Deep Visual Explanations (DVE) also provide a justification for predictions of CNN. Explaining autonomous driving decisions using visual techniques is also primarily motivated by these studies. Particularly, Bojarski et al.'s work [105] was a preliminary explainable vision approach, where the authors proposed 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. [106] 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 [107], 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 and allows to debug the system and understand the rationale for the decisions of a self-driving vehicle.

Kim and Canny [98] use a *causal attention* model on top of the saliency filtering that indicates which input regions actually affect the output (i.e., steering control). Their experiments conducted on the driving datasets - Comma.ai [108], Udacity [109], and Hyundai Center of Excellence in Integrated Vehicle Safety Systems and Control (HCE): this project 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 7). With this approach, the learned framework provides interpretable visualization of a vehicle's actions. As an enhancement of this model, Kim et al. [110] provide textual explanations in their further study. They produce "intelligible explanations" on the decisive actions of a self-driving vehicle using an attention-based video-to-text mechanism and introduce a novel dataset, called Berkeley Deep Drive-X (eXplanation) (BDD-X), that contains annotations for textual explanations and descriptions. Zeng et al.'s [111] 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 that generate interpretable representations as 3D detection of objects, anticipated future trajectories,

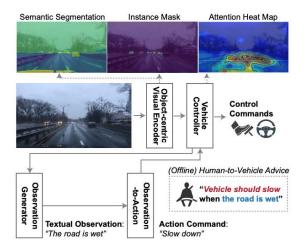


Figure 8: Human advice to a vehicle for appropriate action. Source: [112].

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. [113] propose *object-induced actions* with explanations for predictions of an autonomous car. The authors introduce a new dataset called BDD-OIA, as an extension of the BDD100K dataset [114]; 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. Empirical results of the study performed on the introduced BDD-OIA dataset show that the explainability of the architecture also enhances action-inducing object recognition, resulting in better driving.

In two respective studies, Kim et al., [112, 115] propose an approach that leverages human advice to learn vehicle control (Figure 9). By sensing operational surroundings, the system is able to generate intelligible explanations on the decisive actions (For example, "Stopping because the red signal is on"). 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 vehicle. 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 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 on instead of stopping?" In other words, given the input, counterfactual analysis intends to figure out what are 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. 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. [116] presents 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 elimina-

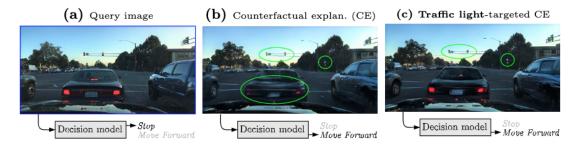


Figure 9: An example of a counterfactual explanation generated by STEEX. Graphics credit: [117].

tion 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. [117] introduces 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. [118], proposes 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 positions of the provided instances.

Overall, we observe that there is a significant focus on perception-based explanations of autonomous driving systems as such explanations provide an opportunity to better understand how accurately a self-driving vehicle senses the operational environment.

5.2. Reinforcement learning and imitation learning-based explanations

Explaining how perceived environmental states are mapped to actions has also recently received attention in the autonomous driving community. In this regard, the field of explainable reinforcement (XRL) is a relatively new and emerging research avenue on explainable AI [119, 120]. Like vision-based explanations, XRL techniques also aim to provide some forms of interpretability on chosen actions of a vehicle either using intrinsically transparent or post-hoc explanations. One of the early works in this context is the Semantic Predictive Control (SPC) framework [121], where the authors propose a data-efficient policy learning approach that predicts future semantic segmentation and provides visual explanations of a learned policy. The framework concatenates multi-scale intermediate features from RGB with tiled actions. The joined modules are then fed into the multi-scale prediction model that predicts future features. Finally, in the last part of the pipeline, the information prediction module inputs the latent feature representation and outputs relevant driving signals alongside the semantic segmentation of the scene. The graphical description of the SPC framework is provided in Figure 10.

Chen et al. [122] introduce a sequential *latent environment model* learned with reinforcement learning and a probabilistic graphical model-based approach that can interpret autonomous cars' actions. They use video cameras and LIDAR images as input in the CARLA simulator. For the purpose of interpretability of actions and explainability of a learned policy, they generate a bird-eye mask. Their model outperforms the used baseline models - DQN, DDPG, TD3, and SAC. Similarly, Wang et al. [123] propose an interpretable end-to-end vision-based motion planning (IVMP) to interpret the underlying actions of an intelligent vehicle. They use semantic maps of bird-view space in order to plan the motion trajectories of an autonomous car. Moreover, the IVMP approach uses an optical flow distillation network that can improve the real-time performance of the network. The experiments conducted on the nuScenes dataset [124] show the superiority of the proposal over modern approaches in semantic map segmentation and imitation of human drivers. In another probabilistic decision-making model, Wang et al. [125] approach lane merging task as a dynamic process and integrate internal states into joint Hidden Markov Model (HMM) and Gaussian Mixture Regression (GMM). The experiments collected on the INTERACTION

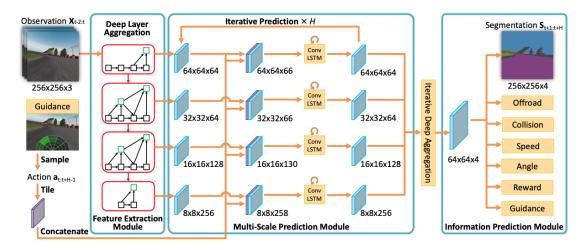


Figure 10: The Semantic Predictive Control (SPC) framework predicts actions and provides visual explanations [121].

dataset [126] demonstrate the efficiency of the proposed technique and show that merging at highway on-ramps can be delineated by three interpretable internal states in terms of the absolute speed of a vehicle while merging.

Recently, Rjoub et al. [127] have shown that federated deep reinforcement learning combined with XAI can lead to trusted autonomous driving. They use a federated learning approach for decision-making and leverage edge computing that enables different devices to train an ML model in a collaborative manner. The model is first developed on the parameter server and further broadcasted to other devices. Then global ML methods intake updates from these devices regularly and the process continues until the model performs well enough on driving tasks.

Finally, within the RL context, a few studies have employed various forms of imitation learning techniques for explainable automated driving. Cultrera et al. [128] present conditional imitation learning with an end-to-end visual attention model, which identifies those parts of images that have a higher influence on predictions. They test their architecture on the CARLA simulator [129] on four tasks - go straight, turn left, turn right, and follow the lane. Their ablation study focused on box type importance and fixed grid analysis to get an attention map on the images shows that integrated imitation learning and attention model enables a car to drive safely and take actions and maneuvers in real-time.

Furthermore, Teng et al. [130] propose hierarchical interpretable imitation learning (HIIL) technique that unifies bird eye view (BEV) mask with the steering angle to perform actions in complex situations as an end-to-end autonomous driving pipeline. They construct their method as a two-phase task: In the first phase, the pre-trained BEV model is used to interpret the driving environment. Then imitation learning takes the latent features of BEV mask from the first phase and combines them with a steering angle acquired through the Pure-Pursuit algorithm. The experiment performed on the CARLA simulator shows that the proposed method enhances the interpretability and robustness of driving in various circumstances. In the most recent work, Renz et al. [131] introduce PlanT, a rigorous imitation-based learning approach that uses transformers for planning. PlanT is able to explain its action decision by recognizing the most-important object in its driving segment, and outperforms state-of-the-art work on CARLA's Longest6 Benchmark by 10 points (See [132] for a visual demonstration). It turns out that the recent success of deep neural network-based attention methods has a huge perspective on the explainability of driving actions.

5.3. Decision tree-based explanations

Being inherently interpretable by design and easier to understand a model's decision, decision-treebased explanations have also been investigated in autonomous driving. Decision trees have been proven to describe the rationale semantically for each prediction made by a CNN architecture [133]. Omeiza et al. [134] 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 employ 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. [135] introduce GRIT (Goal Recognition with Interpretable Trees), a framework that uses decision trees trained from the trajectory data of a self-driving car. The framework, tested on fixed-frame scenarios is proven empirically verifiable for goal recognition using a satisfiability modulo theories (SMT) solver [136].

Cui et al. [137] use Random Forest for the interpretability purpose on the automated car-following task. They use deep reinforcement learning for the decision-making of a 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 [138]. Thus, being computationally more transparent than traditional deep neural network architectures, decision trees can explain behaviors of a variety of autonomous driving tasks with less computation.

5.4. 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 [139] 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 decisive action. The experimental results show that the proposed approach is effective to interpret the safety validation of a car.

Suchan et al. [140] develop an *answer set programming*-based abductive reasoning framework for online sensemaking that is useful 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 the method's explainability and commonsensical value with empirical study collected on the KITTI MOD [141] dataset and the MOT benchmark [142]. Another experimental study leveraging the concept of answer set programming has been carried out by Kothawade et al. [143]: 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.

5.5. User study-based explanations

Finally, except for practical applications, 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. [144] 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 situation awareness

Table 1: Studies on explainable AI-based autonomous driving.

Study	Task	Algorithm(s)	Type of explanation	Target audience
Bojarski et al. [105], 2016	Pixel-based explanations of CNN predictions	CNN	Visual	AVs developers
Kim and Canny [98], 2017	Explaining behavior of a vehicle controller using heat maps	CNN, LSTM	Visual	AVs developers
Kim et al., [110], 2018	Generating textual explanations on a vehicle's control commands	CNN, S2VT, LSTM	Visual and Textual	All groups
Suchan et al., [140], 2019	An answer set programming based abductive reasoning for visual sensemaking	YOLOv3, SSD, Faster R-CNN	Visual	AVs developers
Hofmarcher et al., [106], 2019	Visual scene understanding using semantic segmentation	Enet, SqueezeNet 1.1, ELU	Visual	AVs developers
Pan et al., [106], 2019	Semantic predictive control for explainable policy learning	RL	Visual	AVs developers
Bansal et al., [145], 2019	Imitation learning via a robust policy for a real autonomous car	ChauffeurNet, AgentRNN, PerceptionRNN, Imitation learning	Visual	AVs developers
Zeng et al., [111], 2019	End-to-end interpretable neural motion planner	FaF, IntentNet	Visual	AVs developers
Wiegand et al., [144], 2019	Explaining driving behavior of autonomous cars	User study	Visual	Backup drivers
Cultrera et al., [128], 2020	Explaining autonomous driving by learning end-to-end visual attention	CNN, RL	Visual	AVs developers
Wiegand et al., [146], 2020	Understanding situations that driver needs explanations	User study	Visual	All groups
Xu et al., [113], 2020	Explaining object-induced action decisions for autonomous vehicles	Faster R-CNN	Visual	All groups
Kim et al., [112], 2020	Advisable learning for self-driving vehicles by internalizing observation-to-action rules	Mask R-CNN, LSTM	Visual and Textual	All groups
Corso and Kochenderfer [139], 2020	Interpretable safety validation for autonomous vehicles	Signal temporal logic (STL)	Textual	AVs developers
Li et al., [116], 2021	Risk object identification via causal inference	InceptionResnet-V2, Mask R-CNN, Deep SORT, RoIAlign	Textual	All groups
Schneider et al., [72], 2021	Increasing UX through different feedback modalities	UEQ-S (User study)	Visual, Textual, Audio, Light, Vibration	All groups
Schneider et al., [147], 2021	UX for transparency in autonomous	UEQ-S, AVAM (User study)	Visual, Textual, Light	All groups
Casas et al., [148], 2021	End-to-end model for mapless autonomous driving	CoordConv	Visual and Textual	All groups
Kothawade et al., [143], 2021	Explainable autonomous driving using commonsense reasoning	ASP, s(CASP)	Textual	Road users
Kim et al., [115], 2021	Explainable and advisable model for self-driving cars	DeepLab v3, Mask R-CNN, LSTM	Textual	All groups
Wang et al., [149], 2021	Enhancing automated driving with human foresight	Gaze-based vehicle reference	Visual	Road users
Omeiza et al., [134], 2021	Generating tree-based explanations with and without causal attributions	Tree-based representation / User study	Textual	AVs developers, regulators
Chen et al., [122], 2021	learning	MaxEnt RL, DQN, DDPG, TD3 and SAC	Visual	AVs developers
Wang et al., [123], 2021	Learning interpretable end-to-end vision-based motion planning with optical flow distillation	IVMP, Optical flow	Visual	AVs developers
Albrecht et al., [150], 2021	Interpretable goal-based prediction and planning for autonomous driving	Monte Carlo Tree Search	Textual	Road users
Wang et al., [125], 2021	Uncovering interpretable internal states of merging tasks at highway on-ramps for autonomous driving decision-making	GMR, HMM	Visual	AVs developers
Brewitt et al., [135], 2021	autonomous driving	Goal Recognition with Interpretable Trees, Decision Tree	Visual and Textual	AVs developers
Hanna et al., [151], 2021	Interpretable goal recognition in the presence of occluded factors for autonomous vehicles	Goal and Occluded Factor Inference, Monte Carlo Tree Search	Visual	AVs developers
Mankodiya et al., [138], 2021	venicles	Random Forest, Decision Tree, AdaBoost	Visual	AVs developers
Rjoub et al., [127], 2022	XAI-based federated deep RL for autonomous driving	DQN, DQN-XAI, SHAP	Visual	AVs developers
Madhav and Tyagi, [152], 2022	Explainable navigational intelligence for trustworthy autonomous driving	Grad-CAM, Lime	Visual	AVs developers
Jing et al., [153], 2022	Interpretable action decision making for autonomous driving	Faster R-CNN	Visual and Textual	AVs developers
Jacob et al., [117], 2022	Region-targeted counterfactual explanations	GANs	Visual	AVs developers
Zemni et al., [118], 2022	Object-aware counterfactual explanations	BlobGAN	Visual	AVs developers
Renz et al., [131], 2022	Explainable planning for autonomous driving	BERT, GRU, Imitation Learning	Visual	AVs developers
Teng et al., [130], 2022	Interpretable imitation learning for end-to-end autonomous driving	Bird's Eye View model, Imitation Learning	Visual	AVs developers
Cui et al., [137], 2022	Interpretation framework for autonomous driving	Random Forest, SHAP	Visual	AVs developers
Zhang et al., [154], 2022	Interrelation modeling for explainable automated driving	Faster R-CNN, ResNet-50	Visual	AVs developers

of the drivers. Furthermore, Wiegand et al. [146] investigate situations, in which explanations are needed and methods pertinent 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. [149] propose an approach that enables a human driver to provide *scene forecasting* to an intelligent driving system using a purposeful gaze. They develop a 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 an autonomous car.

Lastly, in their two recent studies, Schneider et al. [72, 147] involve human participants in their empirical studies. They explore the role of explainability-supplied UX in self-driving vehicles. The authors 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. We have summarized the details of the reviewed literature in Table 1.

Based on a high-level overview of all these studies, we see that driving explanations are multi-modal, context-dependent, and task-specific, in general. Moreover, end-to-end learning becomes even more popular for highly-automated decision-making owing to the combination of powerful deep-learning approaches and the availability of advanced sensor devices. As self-driving decisions have a direct impact on road users, the concept of explainability should be accompanied by the safety standards and principles established by transportation regulators. Motivated by this proposition, we can define explainable autonomous driving as follows:

Explainable autonomous driving is automated driving powered by a compendium of AI approaches 1) ensuring an acceptable level of safety for a vehicle's real-time decisions, 2) providing explanations and transparency on the action decisions in critical traffic scenarios, and 3) obeying all traffic rules established by the regulators.

Motivated by this definition and the state-of-the-art works in the above section, we present a general XAI framework for autonomous driving and show the components and processes to achieve safe, regulated, and interpretable automated driving.

6. An XAI framework: integrating autonomous control, explanation, and regulatory compliance

We propose a framework, as a unified approach, in which methods for developing XAI, end-to-end autonomous systems, and regulatory compliance are combined to inform general processes of regulatory principles. Each of the three components has a role in our framework. We have already covered a concise description of such a framework in [15]. We augment that framework with the concepts of simulation and real driving verification, which confirms regulatory compliance. Our integration of autonomous systems with the proposed framework in [15] requires the definition of three constituents of regulated autonomous driving:

1. An end-to-end autonomous control system 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

 $a \in A$.

Within such a formalization, the role of *eeC* is to provide a continuous and complete mapping from the sensed environmental states to relevant driving actions.

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:

$$srC: eeC \cup A$$
.

The safety compliance component is a function that confirms the safety of an eeC system. 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 [45] and [53].

The general principles for acceptable functional safety of road vehicles are defined by the ISO 26262 standard [59]. 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 [155] 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. In this sense, the safety of autonomous driving software should strictly comply with these principles. 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," the latter of which is aligned with our 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 compliant in terms of safety and security. 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., [156]). 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 of automated driving can be considered a "driving school" for self-driving vehicles, as the designed and developed learning software system should be tested rigorously in this phase before such a learning system is deployed to a real vehicle in the physical world.

The alternative, verification, is aligned with our own framework and has significant foundational components established in the discipline of proving software correctness, with a long history (e.g., [157]). The general idea is that an algorithm, for example, an end-to-end autonomous control model or our eeC, can be proven to be correct for all appropriate actions. In our case, the correctness would be confirmed if all the mappings from the environment to action were judged to be secure and safe on real roads.

In our approach, we suggest that regulatory compliance testing systems can be more flexible when they are considered as asking for explanations of intended behavior by an eeC. If a sufficient threshold of explanations is correct and safe, compliance is confirmed. The challenge is in representing the compliance question-asking system and establishing sufficient coverage of alternative behaviors, as is the case in traditional software verification.

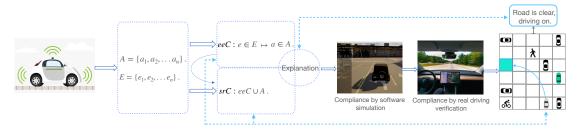


Figure 11: An XAI framework for autonomous driving.

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: 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. It is clear that ideas from software correctness must be coupled with eeC and srC practices in the procedure.

3. An explanation component: This constituent of the framework connects an eeC to a srC. At the highest level, the role of the explanation component is to provide transparency and justification on how an eeC makes a selection from a set of possible actions A. In such a setting, an XAI architecture should be able to provide intelligible explanations for each driving action taken. These explanations can be conveyed in visual, textual, feature importance format or in hybrid, multi-modal ways, as described in the reviewed literature in Section 5. So, XAI-directed autonomous driving should reflect a learned software architecture and regulatory principles at its highest level. In practice, the XAI framework can be realized by directly mapping sensor-based information to a vehicle's control commands using deep learning, reinforcement learning, imitation learning, or their possible combination [158], as attempted by some studies in the literature. A simple graphical illustration of the framework can be seen in Figure 11. While the state-of-the-art end-to-end learning examples reviewed above and more recent works on empirical successes of formal verification of safe driving [159, 160] show a significant advancement in the safety and explainability components individually, achieving safe and explainable autonomous driving in tandem remains unsettled in the state of the art and deserves more attention in future research. What is the best way to achieve the explainability of an AI system in autonomous driving that brings a safety guarantee alongside? Should the research priorities be solely devoted to post-hoc explanations, or one should develop intrinsically interpretable AI architectures? We show with examples that both post-hoc and intrinsic explanations are favorable for autonomous driving. Particularly, one of the seminal studies on the latter topic is Rudin's [161] work: the author logically and empirically proves that instead of interpreting black-box machine learning models in a post-hoc manner, interpretable by-design models should be a preferred choice to design trusted autonomous systems. The author also justifies that the adopted trade-off concept between accuracy and model interpretability is not valid across all domains and domain-specific tasks, such as in computer vision problems. It is possible that an interpretable model can also lead to the same accuracy as the black-box model. From this perspective, we present XAI methods with applicability and canonical examples that validate our framework and show promising steps for future directions.

7. Mind the gap: Future perspective of XAI in autonomous driving

In this section, we propose a set of approaches to guide the pursuit of the goals of XAI for autonomous driving within the principle of the presented framework. In this context, we shed light on the AI approaches that can explain a vehicle's perception system and enable an intelligent driving system to make safe real-time decisions.

7.1. Trustworthy and Debuggable Explainable Vision

7.1.1. Explainable vision through post-hoc explanations

Explainable vision-directed actions for autonomous vehicles are based on how high-level features are used to detect and identify objects. The literature reviewed in Section 5 presents promising attempts to interpret CNN predictions. It is also important that these explanations help to debug, find flaws and improve the existing vision system. We show the importance of vision-directed post-hoc explanations in autonomous driving with two examples. As an alternative to the Trolley problems, International Telecommunication Unit (ITU) recently initiated the "Molly problem" defined as "A young girl called Molly is crossing the road alone and is hit by an unoccupied self-driving vehicle. There are no evewitnesses. What should happen next?" [162]. This is where the value of the post-hoc explanation strategy emerges: The vision system of an automated vehicle could provide a rationale on how it perceived, identified, and distinguished on-road objects and why it continued to drive on and hit the pedestrian. Such explanations are important for transparency in post-accident investigations under regulatory compliance. The second example is the hacking of Tesla's Autopilot by the McAfee Advanced Threat Research team [163]. The team added a sticker to the label of the actual speed limit, 35 mph, and caused the car's heads-up display to perceive it as 85 mph (see Figure 12). The car wrongly accelerated beyond the allowed limit in that traffic area. Even without a careful look at the modified speed limit, humans will not immediately understand why the car sped up in this example. While this is an intentional test performed by humans, similar confusion of the intelligent driving system may be caused by natural phenomena, such as adverse atmospheric conditions. For example, assume that the part of the speed limit sign is covered by muddy rain, and the vehicle's ADAS perceives 3 as 8, increases speed, and potentially causes a traffic accident. So, we see the importance of a post-hoc explanation once again. "Speed limit shows 85 mph, accelerating" would be a simple, timely description to understand the reason for speeding up in case of a regulatory investigation. So, it turns out that the history of descriptive natural language generated along with each relevant action could be helpful to produce reliable post-hoc explanations for critical traffic scenarios. This technique is helpful to debug and improve the existing system as well.

7.1.2. Explainable vision through intrinsic and real-time explanations

As explained in Challange 3 in [161], developing interpretable models that concurrently provide explanations is a promising approach towards achieving a transparent vision system, particularly in object classification tasks. Initial specifications on this perspective use a concept of the *prototype layer* as an addition to a deep neural network: parts of a visual object are decomposed into pieces, and the prototype layer picks out some representative parts of the object during training. When given a new image during testing, the network tries to discover the similarities between those parts of the new image and the ones learned as prototypes in training. In this way, the deep network accumulates evidence from the prototypes and classifies the objects accordingly. Empirical studies of this perspective on classifying handwritten digits, cars, Fashion-MNIST dataset [164], and bird species and cars [165] report nearly the same accuracy as the original black-box models on which they were built. This simple technique delivers a concurrent explanation with a prediction of the neural network and does not require any further post-hoc explanation. In the context of autonomous driving, generating interpretable model-based concurrent explanations can contribute significantly to accident prevention. We support this proposition with a specific example. Assume that a self-driving vehicle has an in-vehicle person (i.e., a backup driver or a passenger). The vehicle provides a control (i.e., stop) button for emergency use. The in-vehicle interface shows there is no human ahead crossing the road and continues driving; but there is an on-road human ahead (i.e., a vision system malfunction). By noticing such an anomaly on time, the in-vehicle person can use the emergency button to slow down and/or stop the car and prevent the accident. This simple example shows that the concept of intrinsic and real-time explanations has potential use in autonomous driving and provides an opportunity for safe navigation of a vehicle.



Figure 12: Extending the middle of 3 by a sticker on the speed limit sign (left side) causes Tesla's ADAS (right side) to read the limit as 85 mph [163]. The red circle on the right side has manually been added to show the causal effect of the performed hacking.

7.2. Explainable Actions Through Model-based Reinforcement Learning

Once an intelligent driving system accurately senses the operational environment, it should map the perceived environment to relevant actions. Autonomous cars' motions can be characterized as sequential decision-making along the trajectory, and so regarded as a Markov Decision Process (MDP). The most commonly explored approaches for autonomous vehicles' learning include three types of MDP: imitation learning, model-free reinforcement learning (RL), and model-based RL, as reviewed in the studies above. Imitation learning intends to mimic the behavior of a human driver; this learning process is computationally expensive as it first must gather real-world driving data as training data [166]. Hence, driving policies obtained under this setting can not be controlled at testing time [167]. Model-free RL algorithms learn by sensing the environment directly and do not have access to the transition dynamics (i.e., prior knowledge) of the explored environment [92]. Such exploration lacks the explainability of a learned policy. A model-free RL agent will explore the driving environment without specific guides if applied in autonomous driving. It may take a long time to learn an optimal driving policy. This problem is directly addressed in model-based RL: An advantage of model-based RL is that an agent learns the model of the environment first, and then adjusts its learning policy according to the dynamics of the environment [168, 169, 170]. This kind of targeted exploration is typically called *planning*, which inherently makes the learning procedure explainable. The idea of planning in RL is vital for proper decision-making and has been investigated in detail in the Dyna architectures [171, 172, 173]. The Dyna and its variation, the linear-Dyna architecture, concurrently learn a world model while learning the optimal policy through interacting with the world. Dyna's planning process creates a predicted future trajectory from an initially provided imaginary state. Based on this structure, model projection generates an optimal action and simultaneously produces a predicted state and a predicted reward. Those last two components can then be visualized, analyzed, and help us understand why the agent prefers selecting a particular action at a particular time step, as a basis for explanation. As each (critical) action of autonomous driving may need an intuitive explanation, the Dyna architecture and model-based RL, in general, can provide tremendous benefits with their explainability functionality.

7.3. Predictive Knowledge as Knowledge Representation

One of the essential steps for safe driving is to *represent* the collected sensor or vision-based knowledge accurately. What would be the best way to represent knowledge acquired from the driving environment?

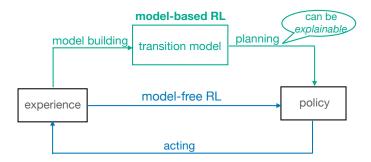


Figure 13: Model-based RL vs. model-free RL from explainability perspective.

For example, assume an autonomous vehicle approaches a four-way intersection. Based on the interaction with the operational surroundings, the intelligent driving system may take the "full stop," "turn left," "turn right," or "go straight" actions. As prior knowledge, these four predictions can be considered as a set of possible actions that the agent (i.e., the car) may choose. This predictive approach to capturing domain knowledge is regarded as predictive knowledge in reinforcement learning, and the idea has acquired growing attention in robotics and autonomous systems research [174, 175, 176, 177]. Moreover, such knowledge is also a reliable basis for explanations of the potential actions taken by an autonomous system. In this context, general value functions (GVFs) provide preliminary techniques for representing predictive knowledge. By definition, GVFs aim to capture long-term predictive summaries about actual observations of an RL agent [177, 178]. For instance, an RL agent in autonomous driving may ask questions and use GVFs to represent the corresponding answers. Examples of questions are "What is the probability I will not face a red signal light in the next intersection?" or "What is the expected time till I arrive at my destination, given my current driving policy?" Experiments performed on robots have demonstrated GVFs' representation value in learning, and their computational accuracy as a proof of concept has been established [177]. Moreover, Kearney et al. [179] have deeply investigated the role of GVFs from an explainable AI perspective, and intuitively analyzed that the question and answer parameters of a GVF can be interpretable. Once an agent inspects its own experience and interprets the decisions to itself, it can also provide an explanatory basis to others through computational estimates. Recent applications of predictive knowledge, formalized as GVFs, on perception problems [180], and learning policies in the real-world autonomous driving [181] setting show the potential benefits of this concept for autonomous driving tasks. Hence, predictive knowledge computationally formalized as GVFs deserves more attention in the ongoing development of representation and explanation for autonomous driving research.

7.4. Incorporating Commonsense Knowledge into Actions: Temporal Questions and Question-driven Software Hierarchy

Another critical aspect of the intelligibility of an autonomous vehicle's actions is strongly related to the proper design of its learning software system. This software system, as an end-to-end framework, should be able to provide a rationale for each action taken at t_n and explain how this particular action leads to the appropriate subsequent action in the t_{n+1} time step. We can infer that *hierarchical software*, corresponding to such principles, is an appropriate structure to support an explainable decision-making system for autonomous driving. Such a structure directly reflects our thoughts while driving, e.g., "Will the traffic light change from green to yellow soon?", "Do the pedestrians ahead intend to cross the road?" or "Why is the car ahead changing its lane?" are just some representative questions that mirror our driving-related considerations while in motion. With this intuition, we can say that the hierarchical software system of an autonomous vehicle can benefit from question-answering functionality. In this sense, leveraging the concepts of visual question answering (VQA) [182] and video question answering (VideoQA) [183] can help understand the behavior of a vehicle and provide causal inference and descriptive rationales on

decisive actions. The state-of-the-art VOA and VideoQA frameworks have had successful applications in various domains such as the medical field [184, 185], advertising [186], and video surveillance [187]. However, this concept has acquired little attention from researchers and practitioners in the autonomous driving community. Understanding upcoming traffic situations and knowing the answers to such questions help us drive carefully and safely. Based upon this context, an explanatory software system can reflect the temporal questions on the temporal actions. As an example, Xiao et al. [188] have recently developed a real-time question-answering system, called NExT-QA, to explain temporal actions from videos in daily common scenes. The framework is able to provide descriptive, temporal, and causal question-answer pairs on the provided video content in various domains and activities. This concept can be applied and bring significant benefits to the autonomous driving domain: The VQA or VideoQA-based driving system may potentially generate explanations as an answer to the asked question (For example, Q: Why is the car stopping at the intersection? - A: Because pedestrians are crossing the road.). In addition, the system can also create *descriptions* that provide information with respect to actions and situations. (For example: Are there other cars in front of the vehicle? - Yes. What was the speed of the car when the emergency brake was used? - 85 km/h, etc.). Hence, driving scene-based question-answering can explain actions and provide descriptive information about *situations*.

From the RL perspective, a suitable approach reflecting temporal states is the concept of options [189]. Options are a generalization of actions in which an RL agent has a policy for taking action with the terminal state. The recently proposed option-critic architecture is motivated by the concept of options [190]. That architecture can learn internal policies and terminal states of options and has proven effective in end-to-end learning of options in the Arcade Learning Environment (ALE). An inherent structure of the option-critic architecture makes it suitable for further development for the learning system of autonomous vehicles. Driving-related questions are often temporal, new questions can be generated for the subsequent actions just after a few seconds. The time sensitivity of driving decisions changes dynamically in real time and exposes the vehicle to different levels of risk. Naturally, actions with lower risks are preferred. Nevertheless, in terms of time and computation, we need to explore efficiently to assess the risk levels associated with the corresponding actions: it is possible that focusing only on increasing RL reward may not lead to desired actions in the long term. As an example, Zhang et al. [191] have shown that not considering risks but only reward as a metric as in traditional RL is not always the perfect decision for an autonomous system; an RL agent may fail to find the optimal policy with such an exploration. In contrast, incorporating different levels of risks with corresponding actions can help discover an optimal policy in environments through different transition and reward dynamics. Accordingly, a well-constructed question hierarchy and evaluation of risk levels concerning appropriate actions can help achieve timely, intuitive, informative, and trustworthy explanations of intelligent vehicles in critical traffic circumstances.

8. Conclusion

We have presented a systematic overview, a design framework, and a future perspective on explainable artificial intelligence approaches for autonomous driving. The key idea is that autonomous vehicles need to achieve regulatory-compliant operational safety and explainability in real-time decisions. Together with a detailed overview of the state of the art in explainable AI-based autonomous driving, our work contributes as a *cause-effect-solution* perspective. We elaborate on the notion of *cause* by identifying current gaps, concerns, and a variety of issues specified while denoting *effect* through the public reluctance on the use of autonomous driving at a broader level. We provide a *solution* through the proposed framework and a set of XAI approaches with examples and applicability for autonomous driving as a future direction. With the provided survey, XAI framework, and future outlook, this paper can benefit the researchers and practitioners to understand the advances in explainability and current limitations of autonomous driving and help enhance this technology to further stages. If the proposed guidelines are implemented properly, we can move a step closer to safer, transparent, publicly approved, and environmentally friendly intelligent vehicles in the near future.

Acknowledgment

We acknowledge support from the Alberta Machine Intelligence Institute (Amii), from the Computing Science Department of the University of Alberta, and the Natural Sciences and Engineering Research Council of Canada (NSERC). Shahin Atakishiyev also acknowledges support from the Ministry of Science and Education of the Republic of Azerbaijan.

References

- [1] S. Singh, Critical reasons for crashes investigated in the national motor vehicle crash causation survey, Tech. rep. (2015).
- [2] K. Bucsuházy, E. Matuchová, R. Zúvala, P. Moravcová, M. Kostíková, R. Mikulec, Human factors contributing to the road traffic accident occurrence, Transportation Research Procedia 45 (2020) 555–561.
- [3] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, A. Weller, Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning, in: IJCAI, 2017.
- [4] O. Pribyl, R. Blokpoel, M. Matowicki, Addressing EU climate targets: Reducing CO2 emissions using cooperative and automated vehicles, Transportation Research Part D: Transport and Environment 86 (2020) 102437.
- [5] R. Abe, Introducing autonomous buses and taxis: Quantifying the potential benefits in japanese transportation systems, Transportation Research Part A: Policy and Practice 126 (2019) 94–113.
- [6] C. Katrakazas, M. Quddus, W.-H. Chen, L. Deka, Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions, Transportation Research Part C: Emerging Technologies 60 (2015) 416–442.
- [7] V. Skrickij, E. Šabanovič, V. Žuraulis, Autonomous road vehicles: recent issues and expectations, IET Intelligent Transport Systems 14 (6) (2020) 471–479.
- [8] B. Friedrich, The effect of autonomous vehicles on traffic, in: Autonomous Driving, Springer, 2016, pp. 317–334.
- [9] Daimler media, Autonomous concept car smart vision EQ fortwo: Welcome to the future of car sharing, (Accessed on October 15, 2021).
- [10] R. Lanctot, et al., Accelerating the future: The economic impact of the emerging passenger economy, Strategy Analytics 5 (2017).
- [11] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
- [12] N. A. Stanton, P. M. Salmon, G. H. Walker, M. Stanton, Models and methods for collision analysis: a comparison study based on the Uber collision with a pedestrian, Safety Science 120 (2019) 117–128.
- [13] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE Access 8 (2020) 58443–58469.
- [14] NTS Board, Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator Mountain View, California. Accessed online on February 10, 2023.

- [15] S. Atakishiyev, M. Salameh, H. Yao, R. Goebel, Towards safe, explainable, and regulated autonomous driving, arXiv preprint arXiv:2111.10518 (2021).
- [16] S. Campbell, N. O'Mahony, L. Krpalcova, D. Riordan, J. Walsh, A. Murphy, C. Ryan, Sensor technology in autonomous vehicles: A review, in: 2018 29th Irish Signals and Systems Conference (ISSC), IEEE, 2018, pp. 1–4.
- [17] D. J. Yeong, G. Velasco-Hernandez, J. Barry, J. Walsh, et al., Sensor and sensor fusion technology in autonomous vehicles: A review, Sensors 21 (6) (2021) 2140.
- [18] J. Shuttleworth, SAE standard news: J3016 automated-driving graphic update, https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic, Accessed online on August 16, 2021 (2019).
- [19] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng, D. Rus, M. H. Ang, Perception, planning, control, and coordination for autonomous vehicles, Machines 5 (1) (2017) 6.
- [20] M. N. Ahangar, Q. Z. Ahmed, F. A. Khan, M. Hafeez, A survey of autonomous vehicles: Enabling communication technologies and challenges, Sensors 21 (3) (2021) 706.
- [21] A. Woo, B. Fidan, W. W. Melek, Localization for autonomous driving, Handbook of Position Location: Theory, Practice, and Advances, Second Edition (2018) 1051–1087.
- [22] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A Survey of Deep Learning Techniques for Autonomous Driving, Journal of Field Robotics 37 (3) (2020) 362–386.
- [23] M. Á. de Miguel, F. García, J. M. Armingol, Improved lidar probabilistic localization for autonomous vehicles using GNSS, Sensors 20 (11) (2020) 3145.
- [24] A. Ziebinski, R. Cupek, D. Grzechca, L. Chruszczyk, Review of advanced driver assistance systems (ADAS), in: AIP Conference Proceedings, Vol. 1906, AIP Publishing LLC, p. 120002.
- [25] Shawn Sinclair, Guide to Cars With Advanced Safety Systems, (Accessed on November 1, 2021).

 URL https://www.consumerreports.org/car-safety/cars-with-advanced-safety-systems-a7292621135/
- [26] J. M. Müller, Comparing technology acceptance for autonomous vehicles, battery electric vehicles, and car sharing—a study across Europe, China, and North America, Sustainability 11 (16) (2019) 4333.
- [27] J. Fleetwood, Public health, ethics, and autonomous vehicles, American Journal of Public Health 107 (4) (2017) 532–537.
- [28] P. Foot, The problem of abortion and the doctrine of the double effect, Oxford review 5 (1967).
- [29] J. Jarvis Thomson, The trolley problem, Yale Law Journal 94 (6) (1985) 5.
- [30] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (7729) (2018) 59–64.
- [31] B. Lundgren, Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles, AI & SOCIETY (2020) 1–11.
- [32] J. Harris, The immoral machine, Cambridge Quarterly of Healthcare Ethics 29 (1) (2020) 71–79.

- [33] S. Sohrabi, H. Khreis, D. Lord, Impacts of autonomous vehicles on public health: a conceptual model and policy recommendations, Sustainable Cities and Society 63 (2020) 102457.
- [34] A. Martinho, N. Herber, M. Kroesen, C. Chorus, Ethical issues in focus by the autonomous vehicles industry, Transport Reviews (2021) 1–22.
- [35] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, Z. Porter, Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective, Artificial Intelligence 279 (2020) 103201.
- [36] L. Collingwood, Privacy implications and liability issues of autonomous vehicles, Information & Communications Technology Law 26 (1) (2017) 32–45.
- [37] A. Taeihagh, H. S. M. Lim, Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks, Transport reviews 39 (1) (2019) 103–128.
- [38] N. Saqib, M. M. Yousuf, M. Rashid, Design and implementation issues in autonomous vehicles-a comparative review, in: 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), IEEE, 2021, pp. 157–162.
- [39] T. Imai, Legal regulation of autonomous driving technology: Current conditions and issues in Japan, IATSS Research 43 (4) (2019) 263–267.
- [40] S. Pettigrew, S. L. Cronin, Stakeholder views on the social issues relating to the introduction of autonomous vehicles, Transport Policy 81 (2019) 64–67.
- [41] A. Nanda, D. Puthal, J. J. Rodrigues, S. A. Kozlov, Internet of autonomous vehicles communications security: overview, issues, and directions, IEEE Wireless Communications 26 (4) (2019) 60–65.
- [42] I. Yaqoob, L. U. Khan, S. A. Kazmi, M. Imran, N. Guizani, C. S. Hong, Autonomous driving cars in smart cities: Recent advances, requirements, and challenges, IEEE Network 34 (1) (2019) 174–181.
- [43] S. Pettigrew, J. D. Nelson, R. Norman, Autonomous vehicles and cycling: Policy implications and management issues, Transportation Research Interdisciplinary Perspectives 7 (2020) 100188.
- [44] M. Koschuch, W. Sebron, Z. Szalay, Á. Török, H. Tschiürtz, I. Wahl, Safety & security in the context of autonomous driving, in: 2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE), IEEE, 2019, pp. 1–7.
- [45] GDPR, Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016, Official Journal of the European Union (2016).
- [46] 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 29-September-2021 (2019).
- [47] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (3-4) (2021) 1–45.
- [48] NACTO, NACTO policy statement on automated vehicles, https://nacto.org/wp-content/uploads/2016/06/NACTO-Policy-Automated-Vehicles-201606. pdf, accessed October 1,2021 (2016).

- [49] NTSB, Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016 (2017).
- [50] Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, author=NTSB Board, note = Accessed online on February 10, 2023.
- [51] National Highway Traffic Safety Administration, Federal automated vehicles policy: Accelerating the next revolution in roadway safety, US Department of Transportation, 2016.
- [52] 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, Accessed online on Apr 5, 2022 (2022).
- [53] Transport Canada, Guidelines for testing automated driving systems in Canada, The Ministry of Transportation of Canada (2021).
- [54] Bundesanzeiger Verlag Board, Act amending the road traffic act and the compulsory insurance act (autonomous driving act) (2021).
- [55] Department for Transport team, Safe use of automated lane keeping system (alks) summary of responses and next steps (2021).
- [56] Austroads, Guidelines for trials of automated vehicles in Australia, National Transport Commission, 2017.
- [57] Advanced information and telecommunications network society, Outline of systematic preparations related to autonomous driving, The Government of Japan, 2017.
- [58] ISO 21448 Committee, Road vehicles Safety of the intended functionality, (Accessed on February 10, 2023). URL https://www.iso.org/standard/70939.html
- [59] ISO 26262 Committee, Road vehicles Functional safety, (Accessed on February 10, 2023). URL https://www.iso.org/standard/68383.html
- [60] Technical Committees, ISO/TC 204 Intelligent transport systems, (Accessed on February 8, 2023). URL https://www.iso.org/committee/54706.html
- [61] CEN/TC 278 Committee, Intelligent Transport Systems, (Accessed on February 8, 2023). URL https://www.itsstandards.eu/
- [62] Apex.AI Blog, An overview of taxonomy, legislation, regulations, and standards for automated mobility, (Accessed on October 8, 2021).
 URL https://www.apex.ai/post/legislation-standards-taxonomy-overview
- [63] U. Ehsan, M. O. Riedl, Human-centered explainable ai: towards a reflective sociotechnical approach, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 449–466.
- [64] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, Y. Li, Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle, in: Designing Interactive Systems Conference 2021, 2021, pp. 1591–1602.

- [65] D. Lewis, Causal explanation (1986).
- [66] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.
- [67] J. Pearl, Causality, Cambridge University Press, 2009.
- [68] 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 13 (2016).
- [69] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, C. Nass, Why did my car just do that? explaining semiautonomous driving actions to improve driver understanding, trust, and performance, International Journal on Interactive Design and Manufacturing (IJIDeM) 9 (4) (2015) 269–275.
- [70] M. Walch, K. Lange, M. Baumann, M. Weber, Autonomous driving: investigating the feasibility of car-driver handover assistance, in: Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications, 2015, pp. 11–18.
- [71] N. Gang, S. Sibi, R. Michon, B. Mok, C. Chafe, W. Ju, Don't be alarmed: Sonifying autonomous vehicle perception to increase situation awareness, in: Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications, 2018, pp. 237–246.
- [72] T. Schneider, S. Ghellal, S. Love, 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.
- [73] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
- [74] D. Holliday, S. Wilson, S. Stumpf, User trust in intelligent systems: A journey over time, in: Proceedings of the 21st International Conference on Intelligent User Interfaces, 2016, pp. 164–168.
- [75] T. Kim, H. Song, How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair, Telematics and Informatics 61 (2021) 101595.
- [76] B. W. Israelsen, 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) 51 (6) (2019) 1–37.
- [77] David Mullen, Mercedes to accept legal responsibility for accidents involving self-driving cars (2022).
- [78] P. Langley, Varieties of explainable agency, in: ICAPS Workshop on Explainable AI Planning (XAIP), 2019.
- [79] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 279–288.
- [80] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

- [81] D. Omeiza, H. Webb, M. Jirotka, L. Kunze, Explanations in Autonomous Driving: A Survey, IEEE Transactions on Intelligent Transportation Systems (2021).
- [82] É. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, Explainability of deep vision-based autonomous driving systems: Review and challenges, International Journal of Computer Vision 130 (10) (2022) 2425–2452.
- [83] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, 2000, pp. 241–250.
- [84] T. R. Roth-Berghofer, Explanations and case-based reasoning: Foundational issues, in: European Conference on Case-Based Reasoning, Springer, 2004, pp. 389–403.
- [85] B. Y. Lim, 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.
- [86] Q. V. Liao, D. Gruen, 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.
- [87] D. Wang, Q. Yang, A. Abdul, 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.
- [88] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
- [89] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
- [90] S. Haykin, R. Lippmann, Neural networks, A Comprehensive Foundation, International Journal of Neural Systems 5 (4) (1994) 363–364.
- [91] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.
- [92] R. S. Sutton, A. G. Barto, Reinforcement learning: An Introduction, 2nd Edition, MIT Press, 2018.
- [93] Z. N. Sunberg, C. J. Ho, 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.
- [94] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
- [95] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating Visual Explanations, in: European Conference on Computer Vision, Springer, 2016, pp. 3–19.
- [96] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
- [97] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, 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.

- [98] J. Kim, 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.
- [99] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).
- [100] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, K.-R. Müller, Interpreting the predictions of complex ml models by layer-wise relevance propagation, arXiv preprint arXiv:1611.08191 (2016).
- [101] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature Communications 10 (1) (2019) 1–8.
- [102] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
- [103] H. K. B. Babiker, R. Goebel, An Introduction to Deep Visual Explanation, 31st Neural Information Processing Systems Conference (NIPS), Long Beach, CA, USA (2017).
- [104] H. Babiker, 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).
- [105] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, K. Zieba, Visual-BackProp: efficient visualization of CNNs, arXiv preprint arXiv:1611.05418 2 (2016).
- [106] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, 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.
- [107] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
- [108] Comma.AI, Public driving dataset, https://github.com/commaai/research, Accessed online on Apr 6, 2022.
- [109] Udacity, Public driving dataset, https://public.roboflow.com/object-detection/self-driving-car, Accessed online on Apr 6, 2022.
- [110] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 563–578.
- [111] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, R. Urtasun, End-to-end interpretable neural motion planner, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8660–8669.
- [112] J. Kim, S. Moon, A. Rohrbach, T. Darrell, 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.
- [113] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, N. Vasconcelos, 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.

- [114] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, 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.
- [115] J. Kim, A. Rohrbach, Z. Akata, S. Moon, T. Misu, Y.-T. Chen, T. Darrell, J. Canny, Toward explainable and advisable model for self-driving cars, Applied AI Letters (2021) e56.
- [116] C. Li, S. H. Chan, 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. 10711–10718.
- [117] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, 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.
- [118] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, OCTET: Object-aware Counterfactual Explanations, arXiv preprint arXiv:2211.12380 (2022).
- [119] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowledge-Based Systems 214 (2021) 106685.
- [120] S. Milani, N. Topin, M. Veloso, F. Fang, A Survey of Explainable Reinforcement Learning, arXiv preprint arXiv:2202.08434 (2022).
- [121] X. Pan, X. Chen, Q. Cai, J. Canny, 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.
- [122] J. Chen, S. E. Li, M. Tomizuka, Interpretable End-to-End Urban Autonomous Driving With Latent Deep Reinforcement Learning, IEEE Transactions on Intelligent Transportation Systems (2021).
- [123] H. Wang, P. Cai, Y. Sun, L. Wang, 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. 13731–13737.
- [124] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuScenes: A Multimodal Dataset for Autonomous Driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11621–11631.
- [125] H. Wang, W. Wang, S. Yuan, 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).
- [126] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, et al., INTERACTION dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps, arXiv preprint arXiv:1910.03088 (2019).
- [127] G. Rjoub, J. Bentahar, 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.
- [128] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, 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.

- [129] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun, Carla: An open urban driving simulator, in: Conference on Robot Learning, PMLR, 2017, pp. 1–16.
- [130] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, X. Hu, Hierarchical Interpretable Imitation Learning for End-to-End Autonomous Driving, IEEE Transactions on Intelligent Vehicles (2022).
- [131] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, A. Geiger, Plant: Explainable planning transformers via object-level representations, in: 6th Annual Conference on Robot Learning, 2022. URL https://openreview.net/forum?id=80vpxjt3vq
- [132] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, A. Geiger, PlanT Project Homepage, https://www.katrinrenz.de/plant/, Accessed online on February 10, 2023.
- [133] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, Interpreting cnns via decision trees, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [134] D. Omeiza, H. Web, M. Jirotka, L. Kunze, Towards accountability: providing intelligible explanations in autonomous driving, in: 2021 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2021, pp. 231–237.
- [135] C. Brewitt, B. Gyevnar, S. Garcin, 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.
- [136] L. De Moura, 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.
- [137] Z. Cui, M. Li, Y. Huang, Y. Wang, 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.
- [138] H. Mankodiya, M. S. Obaidat, R. Gupta, 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.
- [139] A. Corso, 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.
- [140] J. Suchan, M. Bhatt, S. Varadarajan, Out of sight but not out of mind: an answer set programming based online abduction framework for visual sensemaking in autonomous driving, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 1879–1885.
- [141] A. Geiger, P. Lenz, 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.
- [142] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831 (2016).
- [143] S. Kothawade, V. Khandelwal, K. Basu, H. Wang, G. Gupta, Auto-discern: Autonomous driving using common sense reasoning, arXiv preprint arXiv:2110.13606 (2021).

- [144] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, 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.
- [145] M. Bansal, A. Krizhevsky, A. Ogale, ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst, Robotics: Science and Systems (2018).
- [146] G. Wiegand, M. Eiband, M. Haubelt, 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.
- [147] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fülbier, 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.
- [148] S. Casas, A. Sadat, 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. 14403–14412.
- [149] C. Wang, T. H. Weisswange, M. Krueger, 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.
- [150] S. V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, S. Ramamoorthy, Interpretable goal-based prediction and planning for autonomous driving, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 1043–1049.
- [151] J. P. Hanna, A. Rahman, E. Fosong, F. Eiras, M. Dobre, J. Redford, S. Ramamoorthy, S. V. Albrecht, 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.
- [152] A. S. Madhav, 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.
- [153] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, Z. Ding, Inaction: Interpretable action decision making for autonomous driving, in: Proceedings of the 2022 European Conference on Computer Vision, Springer, 2022, pp. 370–387.
- [154] Z. Zhang, R. Tian, R. Sherony, J. Domeyer, Z. Ding, Attention-based interrelation modeling for explainable automated driving, IEEE Transactions on Intelligent Vehicles (2022).
- [155] ISO 26262-6, ISO 26262-6:2018 Road vehicles Functional safety Part 6: Product development at the software level, (Accessed on February 10, 2023).
 URL https://www.iso.org/standard/68388.html
- [156] T. J. et al., Ship collision avoidance and COLREGS compliance using simulation-based control behavior selection with predictive hazard assessment, IEEE Transactions on Intelligent Transportation Systems 17 (12) (2016) 3407–3422.
- [157] S. Hantler, J. King, An Introduction to Proving the Correctness of Programs, ACM Computing Surveys 8 (1976) 331–353. doi:10.1145/356674.356677.

- [158] B. Peng, Q. Sun, S. E. Li, D. Kum, Y. Yin, J. Wei, T. Gu, End-to-end autonomous driving through dueling double deep q-network, Automotive Innovation 4 (2021) 328–337.
- [159] C. Pek, S. Manzinger, M. Koschi, M. Althoff, Using online verification to prevent autonomous vehicles from causing accidents, Nature Machine Intelligence 2 (9) (2020) 518–528.
- [160] H. Krasowski, X. Wang, M. Althoff, Safe reinforcement learning for autonomous lane changing using set-based prediction, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2020, pp. 1–7.
- [161] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.
- [162] ITU Team, The Molly Problem, AI for autonomous and assisted driving(Accessed on November 23, 2021).

 URL https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/
 MollyProblem.aspx
- [163] S. Povolny, S. Trivedi, Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles, McAfee Advanced Threat Research (2020).
- [164] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
- [165] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: Deep learning for interpretable image recognition, Advances in Neural Information Processing Systems 32 (2019) 8930–8941.
- [166] Z. Xu, J. Chen, M. Tomizuka, Guided policy search model-based reinforcement learning for urban autonomous driving, arXiv preprint arXiv:2005.03076 (2020).
- [167] F. Codevilla, M. Müller, A. López, V. Koltun, A. Dosovitskiy, End-to-end driving via conditional imitation learning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 4693–4700.
- [168] H. Yao, C. Szepesvári, Approximate policy iteration with linear action models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 26, 2012.
- [169] H. Yao, C. Szepesvári, B. A. Pires, 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.
- [170] T. M. Moerland, J. Broekens, C. M. Jonker, Model-based reinforcement learning: A survey, arXiv preprint arXiv:2006.16712 (2020).
- [171] R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting, ACM Sigart Bulletin 2 (4) (1991) 160–163.
- [172] R. S. Sutton, C. Szepesvári, A. Geramifard, M. Bowling, Dyna-style planning with linear function approximation and prioritized sweeping, in: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI'08, AUAI Press, Arlington, Virginia, USA, 2008, p. 528–536.

- [173] H. Yao, S. Bhatnagar, D. Diao, R. S. Sutton, C. Szepesvári, Multi-step dyna planning for policy evaluation and control, in: Advances in Neural Information Processing Systems, Vol. 22, 2009.
- [174] G. L. Drescher, Made-up minds: a constructivist approach to artificial intelligence, MIT Press, 1991.
- [175] R. S. Sutton, B. Tanner, Temporal-difference networks, Advances in Neural Information Processing Systems 17 (2004).
- [176] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, D. Precup, Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction, in: The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2, 2011, pp. 761–768.
- [177] A. White, Developing a predictive approach to knowledge, PhD thesis, University of Alberta (2015).
- [178] G. Comanici, D. Precup, A. Barreto, D. K. Toyama, E. Aygün, P. Hamel, S. Vezhnevets, S. Hou, S. Mourad, Knowledge representation for reinforcement learning using general value functions (2018).
- [179] A. Kearney, J. Günther, P. M. Pilarski, Prediction, Knowledge, and Explainability: Examining the Use of General Value Functions in Machine Knowledge, Frontiers in Artificial Intelligence 5 (2022).
- [180] D. Graves, K. Rezaee, S. Scheideman, Perception as prediction using general value functions in autonomous driving applications, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 1202–1209.
- [181] D. Graves, N. M. Nguyen, K. Hassanzadeh, J. Jin, Learning predictive representations in autonomous driving to improve deep reinforcement learning, arXiv preprint arXiv:2006.15110 (2020).
- [182] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.
- [183] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, T.-S. Chua, Video question answering: Datasets, algorithms and challenges, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 6439–6455.
- [184] L. Shi, F. Liu, M. P. Rosen, Deep multimodal learning for medical visual question answering., in: CLEF (working notes), 2019.
- [185] F. Liu, Y. Peng, M. P. Rosen, An effective deep transfer learning and information fusion framework for medical visual question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, Springer, 2019, pp. 238–247.
- [186] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, A. Kovashka, Automatic understanding of image and video advertisements, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1705–1715.
- [187] A. S. Toor, H. Wechsler, M. Nappi, Biometric surveillance using visual question answering, Pattern Recognition Letters 126 (2019) 111–118.

- [188] J. Xiao, X. Shang, A. Yao, T.-S. Chua, NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9777–9786.
- [189] R. S. Sutton, D. Precup, S. Singh, Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning, Artificial Intelligence 112 (1-2) (1999) 181–211.
- [190] P.-L. Bacon, J. Harb, D. Precup, The Option-Critic Architecture, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.
- [191] S. Zhang, B. Mavrin, L. Kong, B. Liu, H. Yao, QUOTA: The Quantile Option Architecture for Reinforcement Learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 5797–5804.