Explainable artificial intelligence for autonomous driving: An overview and guide for future research directions

Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, Randy Goebel

Abstract—Autonomous driving has achieved a significant milestone in research and development over the last decade. There is increasing interest in the field as the deployment of self-driving 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 decision-making 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 a 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 state-of-the-art studies on XAI for autonomous driving. We then propose an XAI framework that considers all the societal and legal requirements for explainability of autonomous driving systems. Finally, as future research directions, we provide a guide to XAI approaches that can improve operational safety and transparency to support public approval of autonomous driving technology by regulators, manufacturers, and all engaged stakeholders.

Index Terms—Explainable artificial intelligence, autonomous driving, intelligent transportation systems, regulatory compliance

I. INTRODUCTION

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) [4], reduced CO₂ emissions [5], diminished transportation costs [6], improved safety [7], [8], and reduced traffic density [9]. In particular, reduced traffic congestion and safety assurance are two significant promises of autonomous vehicles. Intel’s report on the projected benefits of autonomous vehicles 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]. 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 (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 [14]. 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 research makes the following contributions:

- We provide a comprehensive survey of the state-of-the-art XAI approaches for autonomous driving,
- We propose a framework for guiding the development of principles for XAI-based autonomous driving, and
- We propose a guide for the development of XAI techniques for autonomous driving, which conform to the presented framework, with the goal of ensuring public trust and approval.

The rest of the article is structured as follows. In Section II, 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 III. We present a concise overview of common AI approaches powering the real-time decisive actions of autonomous driving in Section IV. In Section V, we provide a comprehensive survey on the cutting edge of explainable AI approaches for autonomous driving. Motivated by the limitations of these studies, we introduce an XAI-based autonomous driving framework that considers safety, public expectations, and regulatory principles for the design and development of a software architecture of intelligent driving systems. Finally, we provide a future perspective of XAI approaches in autonomous driving and sum up our article with Conclusions.

II. BACKGROUND

A. 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 [15], [16]. 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 [17]:

- 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. Perception is the sensing of an operational environment defined as a combination of two tasks: road surface extraction and on-road object detection [18]. 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 [16], [19]. The process of localization enables an autonomous vehicle to accurately determine its position in the sensor model of the world [20], [21]. 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 [22]. 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 [23]. 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 [24]. So, autonomous vehicles can operate on roads without human assistance by fusing information from multi-modal data sources with AI-powered computer vision, decision-making, and control algorithms.

B. 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 [25]. 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 [26] 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 [27], [28] (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 [29], 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:

![Fig. 2: A general structure of AI-driven decision making of an autonomous vehicle in its driving environment.](image)
security concerns [43]. 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 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.

C. AVs regulations and standards

The issues and growing concerns caused by AI systems create the need to scrutinize 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 [44]. 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 [45]. Similarly, Mohseni et al. [46] have tabled a broad description of ethical AI, with 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 [47] 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., [48], [49]) 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 [50]. 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 [51]. The Government of Canada has also recently released their comprehensive federal guidelines on testing and regulations of automated driving systems [52]. 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 the public roads [53]. 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 [54]. Other developed countries such as Australia [55], and Japan [56] have also recently launched initiatives for trials of the 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 [57], which specifies situational awareness standards to maintain operational safety under the “Safety of the Intended Functionality,” and the ISO 26262 [58] 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 [59] 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 [60] 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, involves 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 [61].

III. EXPLANATIONS IN AUTONOMOUS DRIVING

A. 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 explanation of autonomous driving system 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 [62], [63]. From the philosophical point of view, explaining AI decisions can provide descriptive information about the causal history of actions taken [64], [65], particularly in critical situations. 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 [66]. 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 the 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 [67], [68], [69]. 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 [70]. So human-centered design that uses intelligible AI methods is a necessary step for widespread adoption of AVs 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 Ribeiro et al.’s [71] work, “if the users do not trust a model or a prediction, they will not use it.” In their case study, Holliday et al. [72] 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 [73]. Israelson and Ahmed [74] 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 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 the AVs technology. Martinho et al. [33] 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 [44], 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., [34], 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 [75]. Mercedes’s declaration of legal culpability is a significant milestone toward accountability of AVs technology.

**B. 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, sufficiently informative, and intelligible
explanations [76], [77]. Motivated by a target audience definition of [78] and [79], 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.

C. 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 explainee. In fact, explanation construction is one of the major challenges in current explainable AI research. Zablocki et al. [80] 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 [81] and knowledge-intensive case-based reasoning systems [82]. Another empirical approach attempted to derive explanations based on some intelligibility types [83] and used “why,” “why not,” “what if,” and “how to” type explanations for causality filtering. In a recent study, Liao et al. [84] have interviewed twenty user-interface and design practitioners working in different areas of AI to understand users’ explanatory requirements. By doing so, they have attempted to find the gaps in the interviewers’ products and developed a question bank: the authors represent users’ needs as questions so that users may potentially ask about the outcomes produced by an AI system. Overall, the stakeholder needs-based explanation design can be viewed as one of the promising approaches.

Another popular approach to produce explanations is based on using psychological tools from formal theories, according to the literature review of [85]. 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. [79] 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 dependent. 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: [79] 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-vehicles 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.

IV. AI FOR AUTONOMOUS DRIVING

Real-time decisions in autonomous driving are based on environmental perception, processing temporal sequence data, mapping of real-time perception to action. 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 all focus on mapping sensory information to relevant actions. These AI architectures and their enhancements are dominant techniques that have proven their empirical success in the development and deployment of the state-of-the-art AVs.

A. Convolutional neural networks

Convolutional neural networks (CNN) are an AI architecture typically used to process spatial information, such as images and videos [86]. 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, $b$:

$$\theta = [W, b]$$  

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.

**B. 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 [87], [88]. 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 lever 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 $\{x_{t-1}, x_t, x_{t+1}, \ldots\}$ 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

$$h_t = f_H(o_t),$$  

in which

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

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

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

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 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 [89]. 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

---

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

**Fig. 5:** A structure of a typical recurrent neural network.
C. 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 [90]. 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
- $\gamma$ - 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 $\pi^*$ that results in the maximum expected sum of discounted reward:

$$
\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} | s_0 = s \right\}.
$$  \hspace{1cm} (5)

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

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

Basically, the value function is a measure of how good it is for an agent to be in a particular state [90]. 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 [91]. 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 [90]. 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.

V. 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, two types of explanations are common for self-driving vehicles’ actions: visual and textual explanations. At the highest level, visual explanations are about understanding which portions of an image influence a vehicle controller to take particular actions, while textual explanations aim to provide intelligible rationales behind the actions taken by a vehicle, using a natural and understandable language. Our study systematically reviews these investigations that generate some form of explanations for autonomous driving tasks.

As deep neural networks, often in augmented forms as CNNs, power the vision ability of the intelligent vehicles, understanding how CNNs capture real-time image segments that lead to 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. [92] used deconvolution layers to understand the internal representation of CNNs in their seminal work. Hendricks et al. [93] propose a model concentrating on distinguished properties of objects that explain the rationale for the predicted label. Zhou et al.’s [94] saliency map architecture, class activation map (CAM), highlights the discriminative part of an image to predict the label of the image. Moreover, Selvaraju et al. [95] 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, [96], layer-wise relevance propagation [97], [98], and DeepLift [99]. Babiker and Goebel [100], [101] 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: Bojanowski et al. [102] propose a visualization method, called Visual-BackProp, 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. Hofmarcher et al. [103] 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.
Preprocessing

Input images

Attention heat map

1. Encoder: Convolutional Feature Extraction

2. Coarse-Grained Decoder: Visual Attention

3. Fine-Grained Decoder

Output steering angle

1. Encoder: Convolutional Feature Extraction

2. Coarse-Grained Decoder: Visual Attention

3. Fine-Grained Decoder

Input images

Attention heat map

Clustering analysis

Visual Saliency detection and causality check

Output steering angle

Fig. 7: End-to-end learning of steering angle commands from an input image. Source: [104].

[105], 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 [104] 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 [106], Udacity [107], 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. [108] 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.

Cultrera et al. [109] 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 [110] 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.

Zeng et al. [111] propose an architecture that 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, 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 override, input horizons, and training losses on end-to-end learning.

Xu et al. [112] 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 [113]; 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., [114], [115] propose an approach that leverages human advice to learn vehicle control (Figure 8). By sensing operational surroundings, the system is able to generate intelligible explanations on the actions taken (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 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. In this regard, Corso and Kochenderfer [119] present a technique to identify interpretable failures of autonomous

### Table I: The studies on explainable AI-based autonomous driving

| Study                  | Task                                         | Algorithm(s)                  | Type of explanation | Target audience     |
|------------------------|----------------------------------------------|------------------------------|---------------------|---------------------|
| Bojarski et al. [102], 2016 | Pixel-based explanations of CNN predictions | CNN                          | Visual              | AVs developers      |
| Kim and Canny [104], 2017 | Explaining behavior of a vehicle controller using heat maps | CNN, LSTM                    | Visual              | AVs developers      |
| Kim et al., [108], 2018  | Generating textual explanations on a vehicle's control commands | CNN, S2VT, LSTM              | Visual and Textual  | All groups          |
| Suchan et al., [116], 2019 | An answer set programming based abductive reasoning for visual sensemaking | YOLOv3, SSD, FR-CNN          | Visual              | AVs developers      |
| Hofmarcher et al., [103], 2019 | Visual scene understanding using semantic segmentation | Enet, SqueezeNet 1.1, ELU      | Visual              | AVs developers      |
| Zeng et al., [111], 2019  | End-to-end interpretable neural motion planner | FaF, IntentNet                | Visual              | AVs developers      |
| Wiegand et al., [117], 2019 | Explaining driving behavior of autonomous cars | User study                    | Visual              | Backup drivers      |
| Cultrera et al., [109], 2020 | Explaining autonomous driving by learning end-to-end visual attention | CNN, RL                      | Visual              | AVs developers      |
| Wiegand et al., [118], 2020 | Understanding situations that driver needs explanations | User study                    | Visual              | All groups          |
| Xu et al., [112], 2020  | Explaining object-induced action decisions for autonomous vehicles | Faster R-CNN                  | Visual              | All groups          |
| Kim et al., [114], 2020  | Advisable learning for self-driving vehicles by internalizing observation-to-action rules | MaskR-CNN, LSTM              | Visual and Textual  | All groups          |
| Corso and Kochenderfer [119], 2020 | Interpretable safety validation for autonomous vehicles | Signal temporal logic (STL) | Textual             | AVs developers      |
| Kothawade et al., [120], 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, MaskR-CNN, LSTM | Textual             | All groups          |
| Wang et al., [121], 2021  | Enhancing automated driving with human foresight | Gaze-based vehicle reference | Visual              | Road users          |
| Omeiza et al., [122], 2021 | Generating tree-based explanations with and without causal attributions | Tree-based representation / User study | Textual             | AVs developers, regulators |
| Chen et al., [123], 2021  | Interpretable end-to-end autonomous driving with latent deep reinforcement learning | MaxEnt RL, DQN, DDPG, TD3 and SAC | Visual              | AVs developers      |
| Wang et al., [124], 2021  | Learning interpretable end-to-end vision-based motion planning with optical flow distillation | IVMP, Optical flow | Visual              | AVs developers      |
| Albrecht et al., [125], 2021 | Interpretable goal-based prediction and planning for autonomous driving | Monte Carlo Tree Search | Textual             | Road users          |
| Wang et al., [126], 2021  | Uncovering interpretable internal states of merging tasks at highway on-ramps for autonomous driving decision-making | GMR, HMM                      | Visual              | AVs developers      |
| Hanna et al., [127], 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      |
| Brewitt et al., [128], 2021 | Interpretable and verifiable goal recognition with learned decision trees for autonomous driving | Goal with Trees, Recognition Interpretable Decision Tree | Visual and Textual | AVs developers      |
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 a decisive action. The experimental results show that the proposed approach is effective to interpret the safety validation of a car.

Wang et al. [121] 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.

To provide intelligible explanations in self-driving cars, Omeiza et al. [122] propose a *tree-based* representation. They generate scenario-based explanations of different types by mapping observations to actions in accordance with traffic rules. They also employ human evaluation in a variety of driving scenarios and generate Why, Why Not, What If, and What explanations for driving situations that can improve the intelligibility and accountability of automated vehicles.

Chen et al. [123] introduces 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 an 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 outperform the used baseline models - DQN, DDPG, TD3, and SAC. Similarly, Wang et al. [124] 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 birdview space in order to plan motion trajectories of an autonomous car. Moreover, the IVMP approach uses an optical flow distillation network that can improve real-time performance of the network. The experiments conducted on the nuScenes dataset [129] 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. [126] 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 INTER-ACTION dataset [130] demonstrate efficiency of the proposed technique and show that merging at highway on-ramps can be delineated by three interpretable internal states in terms of absolute speed of a vehicle while merging.

Schan et al. [116] 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 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 [131] dataset and the MOT benchmark [132]. Another experimental study leveraging the concept of answer set programming has been carried out by Kothawade et al. [120]: 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 evaluate AUTO-DISCERN’s credibility in real-world scenarios, such as lane changing and right turn operations, from the KITTI dataset.

Finally, except for practical investigations, some studies have attempted to involve users in case studies to understand the effective strategies for explanation generation in autonomous driving tasks. Wiegand et al. [117] 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 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 situation awareness of the drivers. Furthermore, Wiegand et al. [118] investigate situations, in which explanations are needed and methods pertinent for 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 on 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.

Based on a high-level overview of these studies, we see that the main attempts are either developing intrinsically interpretable models or post-hoc explanations for the real-time decisions of autonomous driving. As self-driving decisions have a direct impact on the road users, the concept of explainability should be accompanied by the safety standards and principles established by transportation regulators. As explainability of self-driving is well-aligned with the above
An end-to-end autonomous control system component to achieve safe, regulated, and explainable self-driving. Autonomous driving, and show the components and processes of regulated autonomous driving:

Our integration of autonomous systems with the proposed verification, which confirms regulatory compliance. We augment the framework with the concepts of simulation and real 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 [14]. 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 [14] 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,

   \[ E = \{ e_1, e_2, \ldots, e_n \}, \]

   and a compendium of actions

   \[ A = \{ a_1, a_2, \ldots, a_n \}, \]

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

   \[ C : E \rightarrow A. \]

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 environment to relevant actions. This is a generalized definition: we only focus on autonomous controllers postulating a complete control system for any autonomous system rather than its constituents (we are not immediately interested in sub-controllers for sub-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 [44] and [52].

We also acknowledge that 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., [133]). 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 have safety gaps, especially when consensus on the properties and a scope of a simulation is difficult to achieve.

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., [134]). The general idea is that an algorithm, for example, and 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 environment to action were all judged to be secure and safe.

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 are 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.

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 XAI component: This constituent of the framework connects an $eeC$ to a $srC$. At the highest level, the role of an XAI component is to provide clarity 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 on each driving action taken. So XAI-directed autonomous driving should reflect a learned software architecture and regulatory principles at its highest level. A simple graphical illustration of the aforementioned framework can be seen at Figure 9.

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 [135], [136] 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 [137] work: the author logically and empirically proves that instead of using black-box machine learning models, interpretable models can lead to trusted 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.

VII. 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.

A. Trustworthy and Debuggable Explainable Vision

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 V 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 eye-witnesses. What should happen next?” [138]. 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 [139]. 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 10). 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 ahead shows 85 mph, so 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.

2) Explainable vision through intrinsic explanations: As explained in Challenge 3 in [137], developing interpretable models that concurrently provide explanations is a promising approach towards achieving a transparent vision system, par-
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 [140], and bird species and cars [141] 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 an 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 explanations has potential use in autonomous driving and provides an opportunity for safe navigation of a vehicle.

B. 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. 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 [142]. Hence, driving policies obtained under this setting can not be controlled at testing time [143]. 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 [90]. 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 [144], [145], [146]. 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 [147], [148], [149]. The Dyna architecture concurrently learns 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.

C. 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? 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 [150], [151], [152], [153]. 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, GVF aim to capture long-term predictive summaries about actual observations of an RL agent [153], [154]. For instance, an RL agent in autonomous

Fig. 10: Extending the middle of the 3 by a sticker on the speed limit sign (left figure) causes Tesla’s ADAS (right figure) to read the limit as 85 mph [139].
driving may ask questions and use GVF$s to represent the corresponding answers. Examples of questions are “What is the probability I will not face with a red signal light in the next intersection?” or “What is the expected time till I arrive at my destination, given my current driving policy?” Experiments performed on robots have demonstrated GVF$s’s representation value in learning, and their computational accuracy as a proof of concept has been established [153]. Moreover, recent applications of predictive knowledge, formalized as GVF$s, on perception problems [155], and learning policies in the real-world autonomous driving [156] setting show the potential benefits of this concept for the autonomous driving problems. Hence, predictive knowledge computationally formalized as GVF$s deserves more attention in ongoing development of representation and explanation for autonomous driving research.

D. 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 “Is the car ahead accelerating?” 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. But this concept has acquired little attention from researchers and practitioners in the autonomous driving community. Understanding upcoming traffic situations and knowing answers to such questions help us drive carefully and safely. Based upon this context, an explanatory software system should reflect the temporal questions on the taken temporal actions. Xiao et al. [157] have recently developed a real-time question answering system, called NExT-QA, to explain temporal actions from videos in daily common scenes. This concept can bring significant benefits to autonomous driving tasks: the system can 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, real-time video-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 [158]. 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 [159]. 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 and Yao [160] 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.

VIII. Conclusion

This study presented a systematic overview, a framework, and a future perspective on explainable artificial intelligence approaches for autonomous driving. The key idea is that autonomous vehicles need to provide regulatory compliant operational safety and explainability in real-time decisions. In addition to providing a comprehensive survey of the state of the art in explainable AI-based autonomous driving, our work contributes as a cause-effect-solution perspective. We elaborate on cause through 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 introduced framework and a set of XAI approaches with examples and applicability for autonomous driving as a future direction. This paper, with the provided survey, XAI framework, and future directions
can benefit the researchers and practitioners to understand the state of the art and current limitations of autonomous driving and help advance 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). Shahn Atakishiyev also acknowledges support from the Ministry of 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. Buczak, E. Machtouva, R. Zalgapav, P. Moravcová, M. Kostková, and R. Mikulec, “Human factors contributing to the road traffic accident occurrence,” Transportation Research Procedia, vol. 45, pp. 555–561, 2020.

[3] Daimler media. Autonomous concept car smart vision EQ fortwo: Welcome to the future of car sharing. (Accessed on October 15, 2021).

[4] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller, “Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning,” in IJCAI, 2017.

[5] O. Pribyl, R. Blokpoel, and M. Matowicki, “Addressing EU climate targets: Reducing CO2 emissions using cooperative and automated vehicles,” Transportation Research Part D: Transport and Environment, vol. 86, p. 102437, 2020.

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

[7] C. Kuttrakasaz, M. Qudus, W.-H. Chen, and 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, vol. 60, pp. 416–442, 2015.

[8] V. Skrickij, E. Sabanović, and V. Žuraulis, “Autonomous road vehicles: recent issues and expectations,” IET Intelligent Transport Systems, vol. 14, no. 6, pp. 471–479, 2020.

[9] B. Friedrich, “The effect of autonomous vehicles on traffic,” in Autonomous Driving. Springer, 2016, pp. 317–334.

[10] R. Lanctot, “Accelerating the future: The economic impact of the emerging passenger economy,” Strategy Analytics, vol. 5, 2017.

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

[12] E. Yurtseven, J. Lambert, A. Carballo, and K. Takeda, “A survey of autonomous driving: Common practices and emerging technologies,” IEEE Access, vol. 8, pp. 58443–58469, 2020.

[13] NTS Board, “Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator Mountain View, California,” accessed October, vol. 30, 2020.

[14] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Towards safe, explainable, and regulated autonomous driving,” arXiv preprint arXiv:2111.10518, 2021.

[15] S. Campbell, N. O’Mahony, L. Krapalcova, D. Riordan, J. Walsh, A. Murphy, and C. Ryan, “Sensor technology in autonomous vehicles: A review,” in 2018 29th Irish Signals and Systems Conference (ISSC). IEEE, 2018, pp. 1–4.

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

[17] J. Shuttlenworth, “SAE standard news: J3016 automated-driving graphic update,” https://www.sae.org/news/201901/ sae-updates-j3016-automated-driving-graphic, 2019. Accessed online on August 16, 2021.

[18] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meighjani, Y. H. Eng, D. Rus, and M. H. Ang, “Perception, planning, control, and coordination for autonomous vehicles,” Machines, vol. 5, no. 1, p. 6, 2017.

[19] M. N. Ahangar, Z. Q. Ahmed, F. A. Khan, and M. Hafeez, “A survey of autonomous vehicles: Enabling communication technologies and challenges,” Sensors, vol. 21, no. 3, p. 706, 2021.

[20] A. Woo, B. Fidan, and W. W. Melek, “Localization for autonomous driving,” Handbook of Position Location: Theory, Practice, and Advances, Second Edition, pp. 1051–1087, 2018.

[21] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.

[22] M. A. de Miguel, F. García, and J. M. Armingol, “Improved lidar probabilistic localization for autonomous vehicles using GNSS,” Sensors, vol. 20, no. 11, p. 3145, 2020.

[23] A. Ziebinski, R. Cupek, D. Grzechtha, and L. Chruszczyk, “Review of advanced driver assistance systems (ADAS),” in AIP Conference Proceedings, vol. 1906, no. 1. AIP Publishing LLC, p. 120002.

[24] Shawn Sinclair. Guide to Cars With Advanced Safety Systems. (Accessed on November 1, 2021). [Online]. Available: https://www.consumerreports.org/car-safety/ cars-with-advanced-safety-systems-a729621135/

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

[26] E. Fleetwood, “Public health, ethics, and autonomous vehicles,” American Journal of Public Health, vol. 107, no. 4, pp. 532–537, 2017.

[27] P. Foot, “The problem of abortion and the doctrine of the double effect,” Oxford review, vol. 5, 1967.

[28] J. Jarvis Thomson, “The trolley problem,” Yale Law Journal, vol. 94, no. 6, p. 5, 1985.

[29] A. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, “The moral machine experiment,” Nature, vol. 558, no. 7729, pp. 596–598, 2018.

[30] B. Lundgren, “Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles,” AI & SOCIETY, pp. 1–11, 2020.

[31] J. Harris, “The immoral machine,” Cambridge Quarterly of Healthcare Ethics, vol. 29, no. 1, pp. 71–79, 2020.

[32] S. Sohrabi, H. Khreis, and D. Lord, “Impacts of autonomous vehicles on public health: a conceptual model and policy recommendations,” Sustainable Cities and Society, vol. 53, pp. 102457, 2020.

[33] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, “Ethical issues in focus by the autonomous vehicles industry,” Transport Reviews, pp. 1–22, 2021.

[34] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, “Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical and legal perspective,” Artificial Intelligence, vol. 279, p. 103201, 2020.

[35] L. Collingwood, “Privacy implications and liability issues of autonomous vehicles,” Information & Communications Technology Law, vol. 26, no. 1, pp. 32–45, 2017.

[36] A. Taiebgh and H. S. M. Lim, “Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks,” Transport reviews, vol. 39, no. 1, pp. 103–128, 2019.

[37] N. Saqib, M. M. Youssf, and M. Rashid, “Design and implementation issues in autonomous vehicles—a comparative review,” in 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCARK). IEEE, 2021, pp. 157–162.

[38] T. Imai, “Legal regulation of autonomous driving technology: Current conditions and issues in Japan,” IATSS Research, vol. 43, no. 4, pp. 263–267, 2019.

[39] S. Pettigrew and S. L. Cronin, “Stakeholder views on the social issues relating to the introduction of autonomous vehicles,” Transport Policy, vol. 81, pp. 64–67, 2019.

[40] A. Nanda, D. Puthal, J. J. Rodrigues, and S. A. Kozlov, “Internet of autonomous vehicles communications security: overview, issues, and directions,” IEEE Wireless Communications, vol. 26, no. 4, pp. 60–65, 2019.

[41] I. Yaqoob, U. Khan, S. A. Kazmi, M. Imran, N. Guizani, and C. S. Hong, “Autonomous driving cars in smart cities: Recent advances, requirements, and challenges,” IEEE Network, vol. 34, no. 1, pp. 174–181, 2019.
S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, “Who. Safety & security in the context of autonomous driving,” in 2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE), IEEE, 2019, pp. 1–7.

GDPR, “Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016,” Official Journal of the European Union, 2016.

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, 2019, accessed 29th September 2021.

S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and framework for design and evaluation of explainable AI systems,” ACM Transactions on Interactive Intelligent Systems (TIIS), vol. 11, no. 3–4, pp. 1–45, 2021.

NACTO, “NACTO policy statement on automated vehicles,” https://nacto.org/wp-content/uploads/2016/06/NACTO-Policy-Automated-Vehicles-201606.pdf, 2016, accessed October 1, 2021.

NTSB, “Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016,” 2017.

NTSB Board, “Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, March 23, 2018,” 2020.

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

Department of Transportation, “Occupant protection for vehicles with automated driving systems,” https://www.nhtsa.gov/sites/nhtsa/files/2022-03/Final-Rule-Occupant-Protection-Amendment-Automated-Vehicles.pdf, 2022, Accessed online on Apr 5, 2022.

Transport Canada, “Guidelines for testing automated driving systems in Canada,” The Ministry of Transportation of Canada, 2021.

Bundesanzeiger Verlag Board, “Act amending the road traffic act and the compulsory insurance act (autonomous driving act),” 2021.

The Ministry of Transportation of Canada in Canada,” 2017.

Bundesanzeiger Verlag Board, “Act amending the road traffic act and the compulsory insurance act (autonomous driving act),” 2021.

Department for Transport team, “Safe use of automated lane keeping system (alks) summary of responses and next steps,” 2021.

Department for Transport, “Occupant protection for vehicles with automated driving systems,” https://www.gov.uk/government/publications/the-occupants-protective-measures-guidance.

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

ISO 21448 committee. Road vehicles — Safety of the intended functionality. (Accessed on October 8, 2021). [Online]. Available: https://www.iso.org/standard/57039.html

ISO 26262 committee. Road vehicles — Functional safety. (Accessed on October 8, 2021). [Online]. Available: https://www.iso.org/standard/68383.html

Technical committees. ISO/TC 204 Intelligent transport systems. (Accessed on October 8, 2021). [Online]. Available: https://www.iso.org/committee/54706.html

CEN/TC 278 Committee. Intelligent Transport Systems. (Accessed on October 8, 2021). [Online]. Available: https://www.cen.eu/en/c/359/6/0/1/

Apex.AI Blog. An overview of taxonomy, legislation, regulations, and standards for automated mobility. (Accessed on October 8, 2021). [Online]. Available: https://www.apex.ai/post/legislation-standards-taxonomy-overview

U. Ehsan and M. O. Riedl, “Human-centered explainable ai: towards a reflective sociotechnical approach,” in International Conference on Human-Computer Interaction. Springer, 2020, pp. 449–466.

S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, “Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle,” in Designing Interactive Systems Conference 2021, 2021, pp. 1591–1602.

D. Lewis, “Causal explanation,” 1986.

T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 130, pp. 1–38, 2019.

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

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

M. Walch, K. Lange, M. Baumann, and 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.

N. Gang, S. Sibi, R. Michon, B. Mok, C. Chafe, and 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.

T. Schneider, S. Ghellall, S. Love, and A. R. Gerlicher, “Increasing the user experience in autonomous driving through different feedback modalities,” in 20th International Conference on Intelligent User Interfaces, 2021, pp. 7–10.

M. T. Ribeiro, S. Singh, and 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.

D. Holliday, S. Wilson, and 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.

T. Kim and H. Song, “How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair,” Telematics and Informatics, vol. 61, p. 101595, 2021.

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

David Mullen, “Mercedes to accept legal responsibility for accidents involving self-driving cars,” 2022.

P. Langley, “Varieties of explainable agency,” in ICAPS Workshop on Explainable AI Planning (XAI), 2019.

B. Mittelstadt, C. Russell, and S. Wachter, “Explaining explanations in AI,” in Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 279–288.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamens et al., “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” Information Fusion, vol. 58, pp. 82–115, 2020.

M. T. Ribeiro, S. Singh, and 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.

B. Y. Lim and A. K. Dey, “Assessing demand for intelligibility in vision-based autonomous driving systems: Review and challenges,” arXiv preprint arXiv:2101.05307, 2021.

J. L. Herlocker, J. A. Konstan, and J. Riedl, “Explaining collaborative filtering recommendations,” in Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, 2000, pp. 241–250.

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

S. Haykin and R. Lippmann, “Neural networks, a comprehensive foundation,” International Journal of neural systems, vol. 5, no. 4, pp. 363–364, 1994.
[131] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 3354–3361.

[132] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “Mot16: A benchmark for multi-object tracking,” arXiv preprint arXiv:1603.00801, 2016.

[133] 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, vol. 17, no. 12, pp. 3407–3422, 2016.

[134] S. Hunter and J. King, “An introduction to proving the correctness of programs,” ACM Comput. Surv., vol. 8, pp. 313–353, 09 1976.

[135] C. Yao, S. Manzinger, M. Koschi, and M. Althoff, “Using online verification to prevent autonomous cars from causing accidents,” Nature Machine Intelligence, vol. 2, no. 9, pp. 518–528, 2020.

[136] H. Krasowski, X. Wang, and 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.

[137] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, “This looks like that: Deep learning for interpretable image recognition,” Advances in Neural Information Processing Systems, vol. 32, pp. 8930–8941, 2019.

[138] Z. Xu, J. Chen, and M. Tomizuka, “Guided policy search model-based reinforcement learning for urban autonomous driving,” arXiv preprint arXiv:2005.03076, 2020.

[139] F. Codevilla, M. Müller, A. López, V. Koltun, and 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.

[140] H. Yao and C. Szepesvári, “Approximate policy iteration with linear action models,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[141] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, “This looks like that: Deep learning for interpretable image recognition,” Advances in Neural Information Processing Systems, vol. 32, pp. 8930–8941, 2019.

[142] Z. Xu, J. Chen, and M. Tomizuka, “Guided policy search model-based reinforcement learning for urban autonomous driving,” arXiv preprint arXiv:2005.03076, 2020.

[143] S. Povolny and S. Trivedi, “Model Hacking ADAS to Pave Safer Roads,” PhD thesis, University of Alberta, Canada, 2015.

[144] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup, “Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction,” in The 16th International Conference on Autonomous Agents and Multiagent Systems-Volume 2, 2011, pp. 761–768.

[145] A. White, “Developing a predictive approach to knowledge,” PhD thesis, University of Alberta, 2015.

Shahin Atakishiyev is from Mirzababay village of Gabala, Azerbaijan. He received a BSc Computer Engineering degree from Qafqaz University, Azerbaijan in June 2015, and an MSc Computer Engineering degree with a specialization in Software Engineering and Intelligent Systems from the University of Alberta, Canada in January 2018. Since September 2019, he is pursuing a PhD in Computing Science at the University of Alberta, working at the Explainable Artificial Intelligence (XAI) lab under the supervision of Prof. Randy Goebel. Shahin’s research interests include Explainable Artificial Intelligence, Machine Learning, Reinforcement Learning, and Decision Theory.

Mohammad Salameh received the Ph.D. degree from the University of Alberta under the supervision of Dr. Greg Kondrak and Dr. Colin Cherry, with the main focus on statistical machine translation and sentiment analysis. He is currently a Staff Researcher at Huawei Technologies Canada Company Ltd and leading the neural architecture search group, focusing on gradient-based and reinforcement learning approaches. He co-organized Determining Sentiment Intensity in Tweets (SemEval2016) and Affects in Tweets (SemEval2018) shared tasks.

Hengshuai Yao received the Ph.D. degree in reinforcement learning from the Reinforcement Learning and Artificial Intelligence (RLAI) Laboratory, Department of Computing Science, University of Alberta, in 2014. His thesis is on model-based reinforcement learning with linear function approximation. During his Ph.D. studies, he worked on reinforcement learning theory, algorithms, and web applications. He joined NCSOFT Game Studio, San Francisco, in 2016, where he worked on reinforcement learning in games. He was recently appointed as an Adjunct Professor with the Department of Computing Science, University of Alberta.
R.G. (Randy) Goebel is currently a Professor of Computing Science in the Department of Computing Science at the University of Alberta and Fellow and co-founder of the Alberta Machine Intelligence Institute (Amii). He received the B.Sc. (Computer Science), M.Sc. (Computing Science), and Ph.D. (Computer Science) from the Universities of Regina, Alberta, and British Columbia, respectively.

Professor Goebel’s theoretical work on abduction, hypothetical reasoning, and belief revision is internationally well known; his recent research is focused on the formalization of visualization and explainable artificial intelligence (XAI), especially in applications in autonomous driving, legal reasoning, and precision health. He has worked on optimization, algorithm complexity, systems biology, natural language processing, and automated reasoning.

Randy has previously held faculty appointments at the University of Waterloo, University of Tokyo, Multimedia University (Kuala Lumpur), Hokkaido University (Sapporo), visiting researcher engagements at National Institute of Informatics (Tokyo), DFKI (Germany), and NICTA (now Data61, Australia); he is actively involved in collaborative research projects in Canada, Japan, Germany, France, the UK, and China.