

Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field 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-operating vehicles on roads 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 more reliably without human interventions. 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 these decisions are constructed in order to be regulatory compliant across many jurisdictions. Our study sheds a comprehensive light on developing explainable artificial intelligence (XAI) approaches for autonomous vehicles. In particular, we make the following contributions. First, we provide a thorough overview of the present gaps with respect to explanations in the state-of-the-art autonomous vehicle industry. We then show the taxonomy of explanations and explanation receivers in this field. Thirdly, we propose a framework for an architecture of end-to-end autonomous driving systems and justify the role of XAI in both debugging and regulating such systems. Finally, as future research directions, we provide a field guide on XAI approaches for autonomous driving that can improve operational safety and transparency towards achieving public approval by regulators, manufacturers, and all engaged stakeholders.

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

I. INTRODUCTION

A Survey of the American National Highway Traffic Safety Administration (NHTSA) reports that nearly 94% of road accidents are due to human errors [1]. These human-related mistakes are mainly driver distraction, drunk or otherwise impaired driving, lack of attention, violation of the traffic rules, limited view, and jay-walking pedestrians [2]. The lack of rule

Shahin Atakishiyev is with the Department of Computing Science, University of Alberta, Edmonton, Canada. Email: shahin.atakishiyev@ualberta.ca

Mohammad Salameh is with Huawei Technologies Canada Co., Ltd., Edmonton, Canada. Email: mohammad.salameh@huawei.com

Hengshuai Yao is with the Department of Computing Science, University of Alberta and Huawei Technologies Canada Co., Ltd., Edmonton, Canada. Email: hengshuai@ualberta.ca

Randy Goebel is a Fellow of the Alberta Machine Intelligence Institute, and with the Department of Computing Science, University of Alberta, Edmonton, Canada. Email: rgoebel@ualberta.ca

Correspondence: shahin.atakishiyev@ualberta.ca

Manuscript has been submitted to IEEE for publication.



Fig. 1: A canonical exemplar of explainable AI in autonomous driving: a vehicle provides a natural and intelligible explanation of its real-time decision to bystanders. Graphics adapted and modified from the source [3].

obedience, increasing number of vehicles on roads, and improper road culture have therefore motivated officials, manufacturers, and legislators to make substantial improvements in transportation systems. To prevent traffic accidents and create a better road infrastructure, autonomous vehicles (AVs) have made a substantial leap towards attaining enhanced and affordable transportation systems. The potential benefits of AVs are improved convenience, operational safety 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]. Particularly, reduced traffic congestion and safety assurance are two significant promises of autonomous vehicles. Intel's report on the projected benefits of autonomous vehicles shows 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 people's life from 2035 to 2045, just in the USA [10]. While potential impact and benefits of automated vehicles in everyday life are promising, there is a major societal concern on 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 or functions [11], [12], [13]. As AI approaches provide the foundation for real-time driving actions and operations, engaged consumers and regulatory organizations inspect the intelligent driving system

of a vehicle to comprehend whether inappropriate decisions of a car are the actual cause for 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 help confirm a safe operation. In a recent study, the authors have proposed a framework that describe 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) What are the current issues, gaps, and concerns in the design, implementation, and deployment of autonomous vehicles?
- 2) Why is there a need for explainable artificial intelligence in autonomous driving technology?
- 3) How do industrial priorities inform the choices of research directions?
- 4) 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 overview of the state of the art in the autonomous driving industry, and describe existing multidisciplinary issues and concerns that call for transparent and safe self-operating vehicles,
- We provide a classification of explanations, and the explanation receivers with respect to their identity and background knowledge in the field,
- We propose a framework guiding the principles towards regulated autonomous driving, and
- We propose a field guide on XAI techniques for autonomous driving conforming 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 an overview on currently defined levels of autonomous vehicles. Section III covers the variety of multidisciplinary issues in state-of-the-art autonomous driving. The need for explanations in autonomous driving is given in Section IV. Construction, categories of user-specific explanations, and explanation receivers (i.e., “explainees”) are outlined in Section V. In Section VI, we provide an overview of transportation regulators and law enforcement organizations setting out safety and transparency standards on autonomous vehicles. Section VII shows autonomous vehicle operations and the sensor technologies associated with each of these operations. Subsequently, we provide a framework that integrates the ideas of end-to-end autonomous vehicle control, the role of explanatory AI, and principles of regulatory compliance. Finally, we propose a set of XAI approaches on the development of next-generation autonomous vehicles motivated by the beneficial principles of the framework, and finish the paper with concluding remarks.

II. AN OVERVIEW OF AUTONOMOUS DRIVING

Autonomous vehicles are systems capable of sensing their environment and mapping such sensing to real-time driving

decisions by means of an intelligent driving system. To discern, identify, and distinguish the objects in the operational surroundings, autonomous vehicles use a variety of sensing modalities that capture environmental information for making decisions that control a driving system [15], [16]. Based on a repertoire of operational functionality, two general groups of sensors are currently employed for autonomous vehicles [15], [16], [17]. Internal sensors, also known as *proprioceptive* sensors, detect dynamic measures associated with internal elements of the vehicle such as force, battery voltage, etc. Inertial Measurement Units (IMUs), positioning sensors, encoders and magnetometers are exemplars of proprioceptive sensors. In addition, external state sensors, known as *exteroceptive* sensors, such as video cameras, sonar sensors, light detection and ranging device (LIDAR), radio detection and ranging device (RADAR) capture information on sound amplitude, distance estimations and light intensity from the visuals of a vehicle’s surroundings. These sensors enable an autonomous car to sense its environment, plan a motion trajectory, and make real-time driving decisions. Current autonomous vehicles deployed on road networks have different levels of automation, and are based on the in-vehicle technologies and intelligent capabilities. SAE International (previously known as the Society of Automotive Engineers) has defined six levels of autonomous driving [18]: Level 0 - No automation; Level 1-Driving assistance; Level 2 - Partial automation; Level 3 - Conditional automation; Level 4 - High automation; and Level 5 - Full automation. The details of these automation levels with their descriptions and examples can be viewed in Table I. Based on the classification of the automation levels, we infer that the need for explanations requires an alignment appropriate to the level of automation: as an automation capability of a vehicle increases, the need for explanations of the driving decisions intrinsically grows. In the next section, we review these issues and their impact on the societal views of AVs.

III. CURRENT GAPS, ISSUES, AND CONCERN IN AUTONOMOUS DRIVING

Artificial intelligence approaches, predominately in terms of deep learning algorithms, have brought considerable improvements to many essential components (perception, object detection, planning) of autonomous driving technology. 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 [19]. 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. Fleetwood [20] 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 [21], [22] (hitting a pedestrian on an icy road or a parked car, driving on straightly and hitting five people or changing the direction of the steering wheel and hitting a single person, etc.). Another important aspect of the work is a concern for the potential rights and

TABLE I: SAE International-defined AVs levels with examples: Levels 0-2 are human-supervised driving while Levels 3-5 are highly-automated driving.

AVs level	Description of the level	Example Vehicles
Level 0:	A human driver is responsible for all critical driving tasks such as accelerating/ braking, steering, etc. Driver support may include blindspot warning, automatic emergency braking, and lane departure warning.	Kia Rio 2004, Honda 2005, and other early and before 2000s' vehicles.
Level 1:	The vehicle has automated driving support such as acceleration/braking or steering, but the driver is responsible for all other possible driving operations. Driver support may include lane centering or adaptive cruise control.	The majority of the daily used cars is in Level 1.
Level 2:	Advanced Driving Assistance Systems (ADAS) operations such as steering and acceleration/braking are available. Nevertheless, the human driver should monitor the driving surroundings and take relevant actions when needed.	Toyota Corolla 2018, Nissan Sentra.
Level 3:	The vehicle has more advanced features such as object/obstacle detection and can carry out the most driving operation. However, human supervision is still required to take control of the car at any time.	The 2018 Audi A8 is considered to be the first commercial vehicle that can operate autonomously on geofenced road networks [23].
Level 4:	The vehicle can fulfill all possible driving operations under specific conditions. Use of geofencing is required. Human supervision remains as an option.	Alphabet's Waymo cars claim the Level 4 automation [24].
Level 5:	The vehicle can perform all driving operations in any likely scenario, and no human intervention is required. The Level 5 system can drive the vehicle in all possible road conditions.	There is no Level 5 vehicle in the current automobile industry.

liabilities of passengers sitting inside an autonomous car (for example, by using such a car, does a passenger agree to face with some potential risks; does a passenger have the responsibility to protect other road users if an accident happens?). That paper concludes with *four directions* - clear and cross-disciplinary discussions amongst stakeholders, including driving system's action planning choice of an algorithm; enhancing society's knowledge on the issues and limitations of autonomous driving; confirming society's opinions 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 inevitable, how to crash?) and the Trolley problem, mentioned above. For instance, the Moral Machine experiment [25], a well-known and debatable experiment investigated community's preferences on applied Trolley problems (inevitable accident scenarios with binary outcomes) and state that "these preferences can contribute to developing global, socially acceptable principles for machine ethics." However, further research on this issue condemns this opinion and draws attention to lack of safety [26] and mistakes forcing consideration of such dilemmas [27]. Sohrabi et al. [28] have explored the outcomes of autonomous driving technology on public health in an urban area in their recent study. They found out that this technology can affect public health in thirty two directions: more than half of these pathways (seventeen) are negative, eight are positive, and the remaining ones are unsettled. Another similar study led by Martinho et al. [29] has inspected autonomous vehicles-related ethics present both in scientific literature and various industry reports in California. According to this survey, both the scholarly literature and industry documents highlight safety and cybersecurity as main issues and underline concerns

for moral decision-making algorithms, human-involved control, ethics, design, privacy, accountability, and sustainability, among other concerns. Burton et al. [30] have identified three open problems in the state-of-the-art development of autonomous systems. The first one is the *semantic gap* that emerges when a thorough specification of the system is not provided to manufacturers and designers. Another identified issue is the *responsibility gap*, arising when an accident happens and whether an autonomous system or a human is the cause of this accident remains unresolved. Finally, there is the question of who is responsible for compensating the injured during an accident, which precipitates the third issue: the *liability gap*. That study also shows that the core of these issues is associated with domain complexity, system complexity, and transferring more decision-making functions from humans to autonomous systems. In a relevant case study, other authors relate unpredictable urban environments, driving system complexity, and lack of human override in highly autonomous driving to the enumerated core reasons, respectively. Finally, several studies have attempted to elucidate other related issues such as privacy [31], [32], design and implementation issues [33], legal regulation [34], user concerns [35], IoT challenges [36], integration in smart cities [37], management [38] and security concerns [39]. For more descriptive information, we have illustrated the focal points and calls for action of these studies in Table II. 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 see an immense need for explanations in self-driving cars. Providing explanations of critical operations and decisions can significantly increase the acceptance of autonomous vehicles by the community. In the following section, we discuss the need for explanations in autonomous

driving in more detail.

IV. THE IMPORTANCE OF EXPLANATIONS IN AUTONOMOUS DRIVING

The need for explanations in autonomous driving arises from various issues and concerns mentioned above. First of all, the occurrence of road accidents with the participation of autonomous vehicles is a fundamental practical concern. As careless and hazardous driving can directly impact the safety of passengers and bystanders, people generally require confirmation of safe transportation systems. In addition, understanding the causes of actions or decisions is a natural requirement of the human mind. As stated in Riberio et al's [40] work, "if the users do not trust a model or a prediction, they will not use it." In their case study, Holliday et al. [41] have also empirically shown that providing explanations and perceptible systems significantly increases users' trust in a system. In particular, frequently occurring failures without faithful explanations to stakeholders can significantly damage individual and public trust in intelligent systems. Once trust in an intelligent system is damaged, regaining it can be onerous [42]. Hence, humans naturally want to understand critical decisions of cars in specific scenarios to build trust in using them. If *trustworthiness* exists behind the intelligent decision-makings of a car, it further supports another positive component *transparency*. Once transparency is provided, another requirement, *accountability* is attained, which is related to whether the decisions and actions of the system comply with jurisdictional regulations and standards. Finally, such positive factors develop *fairness* enabling good ethical analysis, support, and causal reasoning of the decisive actions of the autonomous system [43]. These components and their interrelations can be considered essential factors to achieve public approval of automated vehicles.

V. TAXONOMY OF EXPLAINEES AND EXPLANATIONS

A. Classification of Explainees

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 domain knowledge and knowledge characteristic of the explainee is essential to provide pertinent, sufficiently informative and intelligible explanations [44], [45]. Motivated by a target audience definition of [43] and [46], we have redefined the taxonomy of end-users in autonomous driving. Figure 2 provides the identity of such stakeholders and their positions in the corresponding classification.

B. Derivation Methods and Classification of Explanations

As explainees are classified based on their domain knowledge and needs, explanations and their design and evaluation

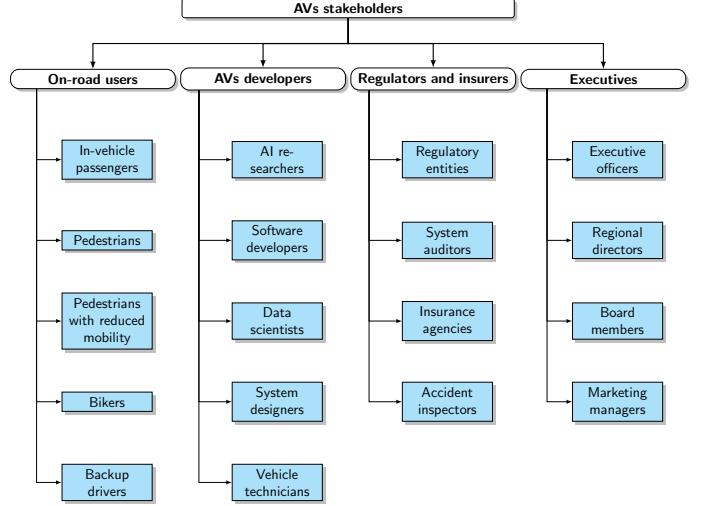


Fig. 2: Taxonomy of the stakeholders in autonomous driving.

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. Generally, 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 [47] and knowledge-intensive case-based reasoning systems [48]. Another empirical work attempted to derive explanations based on some intelligibility types [49] and used "why," "why not," "what if," and "how to" type explanations for causality filtering. In a recent study, Liao et al. [50] 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 AI. So, 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 [51]. Depending on the context and addressee, both explanation derivation methods confirm usefulness. These explanation generation approaches can find alignment in their application in autonomous driving. As autonomous driving involves people with diverse backgrounds in society, relevant XAI design needs inherent adjustments on the context problem.

Like their derivation type, explanations also differ depending on the class they are included. Through their extensive survey, Omeiza et al. [46] 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. In fact, we note that this kind of explanations can be used across many autonomous driving operations.

Explanations based on content type: In this category,

TABLE II: Recent literature focusing on AV-related issues and relevant calls to action.

Study	Investigated AVs issues	Calls to action (if available)
Fletwood et al. [20], 2017	Health and ethical issues	Cross-disciplinary discussions; enriching community's knowledge on the present issues; getting community's input on the present issues; developing justified, faithful, and monitored standards.
Awad et al. [25], 2018	Humans' opinion on decisions in moral dilemmas	Improving society's knowledge on the origin of the ethical principles.
Lundgren [26], 2020	Ethical crashing and safety arguments	Clear definition of "safety"; understanding the outcomes of safety testing and validation.
Harris [27], 2020	Criticizing the principles of Awad et al.'s [25] Moral Machine experiment	If unavoidable, autonomous cars should risk themselves and passengers inside and not bystanders outside the vehicle.
Sohrabi et al. [28], 2020	Negative and positive impacts of AVs on public health	Encouraging ride-sharing; establishing policies to control transportation requirements; motivating humans for electric vehicles; switching to AVs gradually.
Martinho et al. [29], 2021	Focal points of scientific papers vs industry reports	Getting opinions of more humans or focus groups in AV industry; participation of scientists in manufacturing of AVs for a long time.
Burton et al. [30], 2020	Engineering, ethical, and legal issues	Reflective equilibrium on safety, surveillance, and dynamic assurance, clearly defined soft law, regulation and governance.
Collingwood [31], 2017	Information privacy and liability issues	Quick and decisive actions to be taken by law enforcement.
Taeihagh and Lim [32], 2019	Five interconnected issues (safety, liability, cybersecurity privacy and industry risks)	Not available.
Saqib et al. [33], 2021	Design and implementation issues	Not available.
Imai [34], 2019	Legal regulation of AVs in Japan	Well-grouped technical specifications; vehicles for particular geographical areas; keeping the concept of driver (not necessarily inside the car, but somehow to control the car in specific scenarios); preparing for unforeseen circumstances.
Pettigrew et al. [35], 2019	User opinions on the social issues of AVs in Australia	The use of publicly shared AVs; calling for strong attention on the data and engineering perspective of AVs.
Nanda et al. [36], 2019	Communication security of internet of AVs	Enhanced energy efficiency, connectivity, security, safety, and data processing approaches.
Yaqoob et al. [37], 2019	Integration of AVs to smart cities	Improved security; radar interference management; diverse vehicular networks; robust AI algorithms; improving network performance by edge computing.
Pettigrew et al. [38], 2020	Policies and management issues on inter-relationship between bicycles and AVs	Developing public policies to provide safe presence of bikes and AVs on roads.
Koschuch et al. [39], 2019	Cyber-physical security and safety issues and their causal chains	Extending the failure of combined security and safety chain into the causal and consequence perspective, and built a more robust chain to test and certify autonomous cars.

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.

Explanations depending on a model: Here, explanations are distinguished by being either model agnostic (i.e., universal) or model specific. For instance, some autonomous driving operations can be condition-specific and some can be general, regardless of the driving conditions.

Explanations based on a system type: This category attempts to capture the properties of the operational system: [46] provides two natures of explanations as either *data-driven* (i.e., explaining the outcome of 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 a follow-up question to further understand a provided explanation. This feature brings interactivity into the explanation framework.

Explanations with concrete scope: This category captures the feasibility and range of explanations that system can generate by being either *local* or *global*. Local explanations are limited to explanations on some or subset of all possible actions. Global explanations, on the other hand, are capable of explaining all potential actions a system can take.

VI. AVS' REGULATIONS AND STANDARDS

The issues and growing concerns caused by the use of AI systems have necessitated a scrutinized 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 [52]. 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, all to be applied in AI-based product research and development [53]. Similarly, [54] 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 vehicle in confirming safety of autonomous systems and in support of 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. The National Highway Traffic Safety Administration (NHTSA) of the US Department of Transportation has launched a federal guideline on automated vehicle policy to improve traffic safety [55]. NACTO's (National Association of City Transportation Officials) statement on automated vehicles [56] 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, cover promises and issues of autonomous vehicles, association of this technology to law, liability issues. Their principles also provide thorough guidance for public regulators to investigate transportation accidents and makes safety recommendations (e.g., [57], [58]) to the regulators and manufacturers of autonomous vehicles such as NHTSA, SAE International, Tesla, and Apple. The Government of Canada has also recently released their comprehensive federal guidelines on testing and regulations of automated driving systems [59]. Their documentation provides detailed information and a 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 [60]. The UK has also advanced its interests toward regulated and safe autonomous driving and hands-free driving and is expected to be legally allowed by the end of 2021 [61]. Other developed countries such as Australia [62], and Japan [63] 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 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 [64], which specifies situational awareness standards to maintain operational safety under the "Safety of the Intended Functionality," and the ISO 26262 [65] 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 [66] is the primary standard that provides a comprehensive guide on the overall system and infrastructure aspects of intelligent transportation systems (ITS), supporting 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 [67] 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 [68].

VII. AUTONOMOUS VEHICLE OPERATIONS AND UNDERLYING EXPLANATIONS

The real-time decision-making and actions of autonomous vehicles involve several interconnected operational stages. These operations are commonly categorized as *sensing*, *localization*, *planning*, and *control*. In the following subsections, we provide an overview of these operational components and relevant explanation methods associated with them.

A. Sensing

Accurate sensing of operational surroundings is one of the primary functional requirements for the safe driving of an autonomous vehicle. In general, perception in autonomous driving is defined as a combination of two tasks - road surface extraction and on-road object detection [69]. Information for the purpose of perception can be obtained from multi-modal data sources, including LIDAR, RADAR, cameras, and ultrasonic sensors [16], [70]. These devices differ by the type of information they capture and their sensitivity to the environment (e.g., weather, light, obstacles).

LIDAR generates a three-dimensional map of the operational environment for positioning and object detection of autonomous vehicles [16], [69], [70]. It enables a car to "see" its surroundings and, potentially, navigate safely. LIDAR sends light pulses and receives reflected signals from objects over a 360-degree scan. The returned signals are typically noisy and not ideal; some points can be sparse or missing, and inappropriate processing may lead to inferences that result in unsafe actions. To cope with this problem, captured 3D points are represented by points clouds, features, and grids which become input for further processing. The LIDAR systems

used in automated driving have 905 and 1550 nm wavelength; while LIDAR using 905 nm spectrum are better in adverse (rainy, dusty, foggy, etc.) weather conditions and are energy efficient, they are harmful to an eye's retina [70]. Hence, LIDAR with a 1550 nm spectrum is preferred in modern autonomous vehicles, to diminish the potential negative effect on humans' visual systems. LIDAR devices can detect objects up to a circumference of 200 meters of their surroundings and are generally classified as 1D, 2D, and 3D LIDAR systems. 1D LIDARs are limited to measuring x-distance from an object. 2D LIDAR provides reflections on x and y-dimensions. 3D LIDAR sensors provide a three-dimensional mapping of an object (i.e., profiling of objects with their three dimensions to enable a car to understand the object's full shape) [70]. As both surface and on-road objects are crucial for the safe motion of autonomous vehicles, the 3D spinning LIDAR sensor is more practical and efficient, as it can sense a wider view in terms of distance and depth and provide more accurate signals.

RADAR is another active sensory device that scans the environment using radio waves instead of light. It can capture positions of vehicles and other on-road objects located within 5 to 200 meters [70], [71], [72]. Currently, autonomous vehicles use four different frequencies - 24, 60, 77, and 79 GHz, also called millimeter wave RADAR, to detect the objects in the coverage area. Modern RADARs deployed in autonomous vehicles can also detect the speed and directions of moving objects. Unlike LIDAR, a RADAR has a wider field of view, allowing it to better detect obstacles. It is also less sensitive to atmospheric conditions. But due to its smaller wavelengths, it can not recognize objects as well as LIDAR. That said, RADAR sensors have a major downside of false detection of objects; sometimes signals do not provide accurate information in discerning static objects, such as road markings [16]. Therefore, relying solely on RADAR itself is not appropriate for acquiring sensory knowledge. Autonomous vehicles researchers and practitioners often use combinations of sensors use a variety of information fusion methods with these sources to support reliable and precise decisions for autonomous cars.

To mimic human vision for road visual recognition, *video cameras* can provide high-quality images that state-of-the-art computer vision algorithms can process. They enable object recognition in the surroundings including relatively accurate reading of road signs: these features give them an advantage over LIDAR. But a limitation is the inability to provide the objects' distance and location as well as LIDAR. From the perspective of sensing ability and wavelengths, cameras are classified as either *visible* (VIS) and *infrared* (IR) cameras [71]. VIS cameras detect wavelengths in the range 400 nm–780 nm, which is similar to the visible spectrum of a human eye. Because of affordable cost and capability to provide high resolution imaging, these cameras are used in many perception tasks such as object recognition, blind spot detection and road sign detection. But VIS cameras are sensitive to adverse weather conditions. For example, rain, snow, or fog significantly reduces their vision quality. Hence, video cameras are typically combined with other sensors to collect data from the operational surroundings. IR cameras are sensitive to infrared

wavelengths ranging from 780 nm to 1mm. The primary use cases of such cameras are the circumstances where the level of lighting is high (for example, in instances where vehicle driving in bright sun light or when exiting a dark tunnel) [73]. *Ultrasonic sensors*, are another kind of sensing device, which use sonic waves in the range of 20 kHz to 40 kHz to calculate the distance to an object [70], [71]. The operating range of ultrasonic sensors in terms of distance is usually limited to less than 3 meters; thus, it can only be useful to detect close objects at low speed. Similar to other sensory devices, ultrasonic sensors also have pros and cons. The disadvantages include inefficiency in a blind zone and a false positive signal due to bouncing. The positive aspects, on the contrary, include good performance in adverse (e.g., dusty, foggy, humid, rainy) atmospheric conditions and independence on materials from which signals are reflected. Thus, ultrasonic sensor data fused with other sensory devices can provide improved sensing reliability for an autonomous car.

Just as accurate sensing of the environment is an essential requirement of autonomous driving, we suggest that providing underlying explanations of autonomous action decisions is also vital to understanding scene navigation and driving behavior in general, particularly in critical scenarios. Accordingly, several attempts have been made to provide explanatory methods in the perception task of autonomous vehicles. Some research has sought post-hoc explanations by introspective textual descriptions using visual attention [74], some have focused on explanations as descriptions of human behavior-involving causal reasoning [75], [76], and others have focused on object-induced action decisions [77]. Another popular approach to sensing-based explanation generation has been to interpret the outcomes of convolutional neural networks (CNNs). The main idea behind such a method is to measure and display the gradients backpropagated from the output layer of the neural network to the input. Examples of gradient-based explanation methods include Class Activation Map (CAM) [78], its augmented variations such as Guided Grad-CAM [79], Grad-CAM [80], Grad CAM ++ [81], Smooth Grad CAM++ [82], and backpropagation-based methods such as guided-backpropagation [83], layer-wise relevance propagation [84], [85], VisualBackProp [86] and DeepLift [87], [88]. Moreover, Babiker and Goebel [89], [90] have shown that heuristics-based Deep Visual Explanations (DVE) provide a reasonable rationale for predictions of deep CNNs. A comprehensive survey on explainable self-driving systems based on computer vision can be looked at in Zablocki et al.'s recent work [91].

B. Localization

Understanding a vehicle's surroundings depends heavily on the repertoire of sensing devices. The perception from multiple sensors and fusion of their captured data is what informs an autonomous system's model of its surroundings. But sensors are prone to lack of accuracy and latency, thus causing a discrepancy between an autonomous vehicle's high-definition (HD) map model and the sensors' signals. The idea of localization is to enable an autonomous vehicle to accurately locate its position in the sensor model of the world [17], [92]. This is done by comparing the location of reflected

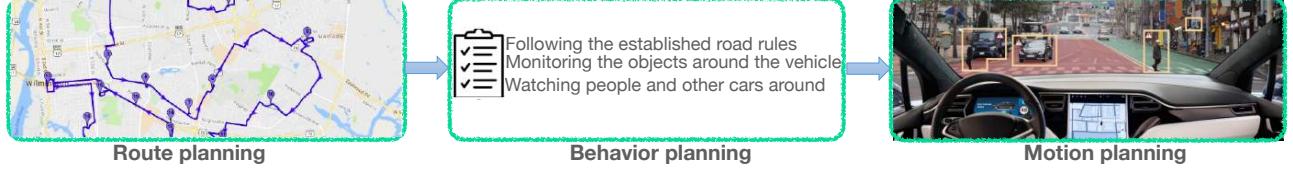


Fig. 3: Motion planning of an autonomous vehicle: The car uses prior knowledge obtained from route and behavior planning and makes appropriate decisions. This figure is based on the graphics in the sources [93], [94].

objects to the HD maps to enable precise positioning. Currently, an effective way to get a position of autonomous vehicles is to use satellite and 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 [95]. The GNSS consists of several satellites rotating approximately 20,000 km above the Earth, which provide signals to the receiver of a vehicle on its dynamic positioning. GPS, in particular, comprises 24 satellites and can provide signals from up to 12 satellites to an arbitrary receiver located on the surface of the Earth [71]. This satellite system delivers a signal based on the time of flight (ToF) between the satellite and the receiver in a vehicle. While the satellite system ensures precise positioning of vehicles, some other sensors provide helper functions. In particular, GPS signals may not be transmitted in places like underground tunnels and canyons or in built infrastructure like snow and avalanche shelters. Such circumstances require alternative methods or exploitation of other sensor technologies. Inertial Measurement Units (IMUs), as one of the helper sensors, are used for control, safe navigation, and car direction. But they have a disadvantage of accumulating errors over time, eventually leading to accuracy drift [15]. Hence, inertial sensors must be combined with GPS to regularly correct IMU-related errors. So, precise localization of autonomous vehicles can also be attained with proper signal transmission and interchange of various sensors and navigational systems.

As real-time decision-making of an automated vehicle requires accurate sensing of on-road position, it is also pivotal to understanding how a position of a vehicle from different navigational systems and sensors is obtained. That is why localization also require an explanation capability. The need to understand the induced position of a self-driving vehicle emerges, especially when the signals coming from GPS or other sensors are imprecise. Such unreliable communication channels may thus force an autonomous car to make incorrect high-stakes decisions. Hence, debugging the navigation systems and related sensors can help prevent inaccurate signals and provide reliable communication for a self-driving car's correct longitudinal and lateral positioning.

C. Planning

Once real-time accurate environmental perception and localization are achieved, an autonomous car plans its trajectory from the initial point to the final destination. Motion planning is a complex operation that needs to consider several factors, such as an interaction with other vehicles, people to be poten-

tially met on a trajectory (i.e., pedestrians, bikers, and other bystanders), dynamics of the environment, and availability of the navigation-assisting resources and infrastructure. The studies on motion planning of autonomous vehicles use a variety of different terms for relevant components comprising the entire planning process. But overall, the planning of a self-driving vehicle can be thought of as a hierarchical process integrating three essential constituents [96]. The first one is *route planning* is where a vehicle chooses its route in the road network from its starting point to the final destination. Route planning can be a challenging problem: the search space to be explored is potentially large. Traditional path planning methods such as Dijkstra's algorithm [97] or A* algorithms [98] may be too inefficient to meet the demand for real time navigation decisions, especially given that the road networks are potentially exponential, but are also dynamic. Consequently, the transportation science and research community has sought better and more reliable algorithmic approaches to overcome the limitations of the classical route planning methods. For example, Goldberg and Harrelson [99] have extended an A* method with Euclidean bounds and empirically demonstrated the effectiveness of their proposed method, specifically on the path planning problem for road networks. In another study, Geisberger et al. [100] propose a concept of contraction hierarchies for finding fast path routing. Their method exploits the hierarchical structure of a road network and uses an adaptive bidirectional Dijkstra algorithm to find the shortest distance between the two points. The empirical results of the proposed approach have been effective in complicated scenarios like traffic congestion. Extensive surveys on route planning of autonomous vehicles with relevant algorithms can be seen at [7], [69], [96], [101]. As intelligent driving system determine a vehicle's route planning, the next step is to complete *behavior planning* that complies with established traffic rules and interacts safely with road participants. Examples of behavior planning include components like stopping at a red light, driving on in a green light, and yielding to pedestrians and other vehicles as appropriate [69], [96]. Behavioral planning is a dynamic task where an agent can update its preferences and actions according to the present conditions. This kind of planning deals with many uncertainties in traffic (e.g., changes in weather conditions, the apparent goals of other traffic participants, and (un)expected rule violations). Therefore, coming up with robust algorithms on traffic flow forecasting and intent prediction is an actual research area in transportation planning, mostly using real time data-driven approaches. Several attempts in the DARPA urban challenge [102] have exploited some form of finite-

state automata (FSA) methods [103], [104], [105], [106]. However, FSA-based approaches can typically not handle complicated traffic situations properly beyond predefined rules. Some further research in this vein employ linear-temporal logic (LTL) as a specification language to attain rule-compliant navigation in an urban environment [107], [108]. On the contrary, more recent works on identifying what is referred to as “intent prediction” have leveraged other state-of-the-art AI approaches. For example, Fernando et al. [109] propose deep inverse-reinforcement learning to accurately predict the behavior of other on-road vehicles and participants for autonomous vehicles and provide both quantitative and qualitative assessments of this method. Another study on the dynamics of planning by Saleh et al. [110] introduces what they call a Spatio-temporal DenseNet [111], which is a model-based framework to predict actions of pedestrians using RGB camera, and reports a average precision of 85%. In further work, Liu et al. [112] propose a graph convolution-based framework to discover pedestrians’ intent of crossing or not crossing the street based on a real-time scene and achieve nearly 80% accuracy. Moreover, Pal et al. [113] use the concept of scene semantics to understand both a vehicle’s and humans’ intention in traffic scenarios and correctly identify 49 cases out of the possible 56 cases. A broad survey of identifying on-road behavioral intentions for safe autonomous driving can be viewed at [114].

Finally, as an intelligent driving system accomplishes some form of route and behavior planning, an autonomous vehicle incorporates this prior knowledge into *motion planning*, the highest level of planning, which supports the choice of safe decisions in the machine’s motion trajectory and enables reliable control and navigation throughout the journey (Figure 3). As planning is a vital component of the real-time decision-making of an autonomous car, understanding planning strategies can improve the safety of a car in complicated traffic scenarios. Hence, explainable planning is required to justify specific plans favoured for a particular task. This kind of explanatory support is an important task towards attaining explainable AI: users of different backgrounds want to understand what logical processes eventually lead to specific action decisions of an autonomous agent. One of the early examples of explainable planning is Bidot et al.’s [115] verbal plan explanations, which provide domain-agnostic explanations of plans based on verbal inputs given by humans. Fox et al. [116] provide a detailed study on how to explain planners and what can be explained on planning, using six concrete questions. Motivated by this work, Borgo et al. [117] propose a framework called XAI-PLAN. This framework enables users to probe alternative questions by suggesting different actions on an original plan. Once users provide their opinions, the XAI-PLAN framework then provides a justification for the initial plan. In this way, user interaction can lead to improved choices of alternative plans. In a similar fashion, a framework called WHY-PLAN explains why an autonomous robot makes specific decisions in its navigation [118]. Furthermore, Krarup et al. [119] provide a list of compiled questions to achieve domain-independent explainable planning. Finally, a thorough overview of prominent methods for explainable planning in

AI systems are included in Chakraborti et al.’s survey [120].

D. Control

Control of an autonomous vehicle requires proper execution of planned motions. This function is mainly done by feedback controllers. A feedback controller chooses relevant inputs from an actuator, fulfills the relevant motion, and corrects tracking errors emitted from the actuation variables. In general, the whole control of autonomous terrestrial vehicles consists of two essential components: longitudinal and lateral control. Longitudinal control is concerned with the longitudinal dynamics of the vehicle, such as the vehicle’s velocity, acceleration, the distance to the cars behind, and the distance to the cars in front. Sensor signals are used to measure the relevant indications. For instance, the velocity and acceleration of a vehicle is calculated by speed sensors and accelerometers, while the distance from the preceding and in front vehicles are gauged using a combination of RADAR, LIDAR, ultrasonic sensors, and cameras (as described above). The second component, lateral control, adjusts the vehicle’s sidelong motion, such as lane change and similar maneuvers. For lateral dynamics, the controller takes the vehicle’s steering as an input variable as steering regulates the steering angle and driving direction of the vehicle. Note that the full spectrum of such control can be more complex than just standard vehicle steering (e.g., 4-wheel steering), but we concentrate here on conventional steering components. One of the most successful and prevalent control methods is called model predictive control (MPC). Such a controller predicts motion for a short time interval, known as the prediction horizon, where the main goal is to calculate the optimal solution for the scope of the prediction horizon. MPC has currently been successfully applied in several control applications in the context of autonomous driving, such as lane-keeping [121], combined braking and steering [122], and navigating dynamically in adverse external conditions [123]. Overall, the process of controlling an autonomous vehicle is typically carried out through embedded control software called *advanced driver-assistance systems* (ADAS). These systems interact with the sensors of an environment and assists the vehicle in controlling its trajectory along the journey [124]. The canonical examples of ADAS include adaptive cruise control, anti-lock braking system, collision avoidance systems, forward collision warning, and lane departure warning systems [125]. One of the most recent and advanced ADAS systems has been developed by Mobileye, an Intel company that includes both *passive* and *active* control systems in their tool suite [126]. The passive control system, such as lane departure warning functionality, is in charge of notifying the driver to avoid potentially hazardous scenarios ahead. On the other hand, an active control system takes relevant action to prevent potential danger while also providing the appropriate information to the driver.

As vehicle control eventually reflects high-level decisions of driving system, the users may require timely explanation of the rationale for real-time autonomous action choices. This demand brings the essence of explainability into the control system of automated vehicles. As suggested in [46], in-vehicle

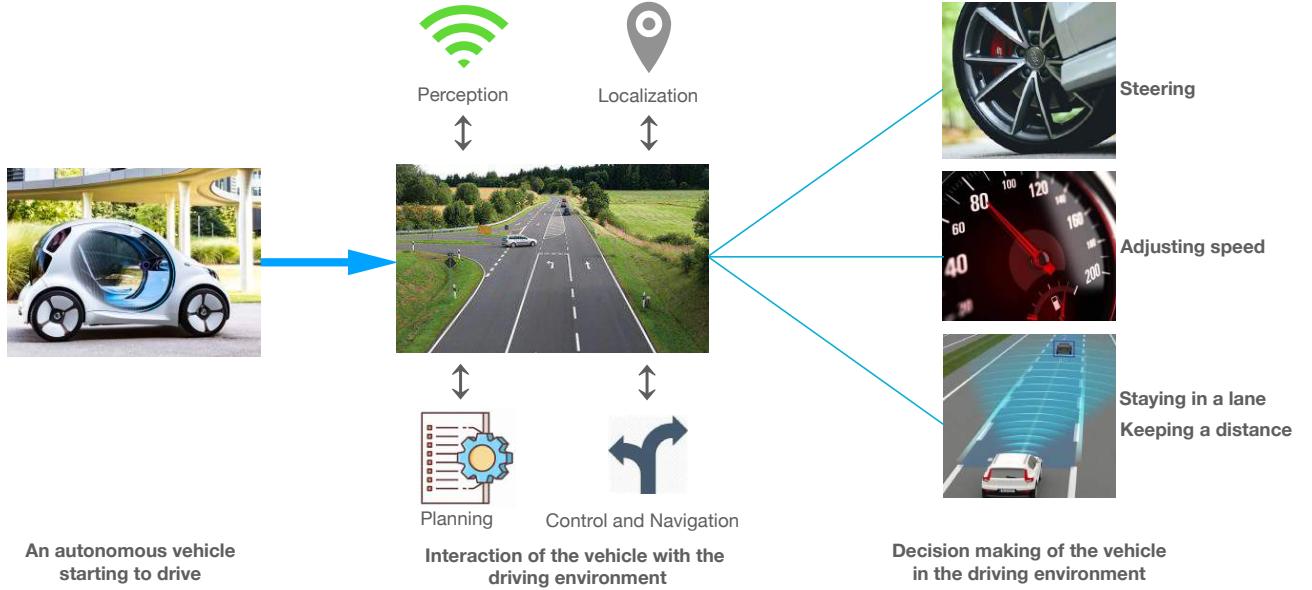


Fig. 4: A general structure of the decision making of an autonomous vehicle in its driving environment.

interfaces, dashboard panels, and other user-friendly functionalities help users ask the “why” questions (e.g., “Why did you stop at the right side?”), or contrastive questions (e.g., “Why did you select this route instead of the other route?”), counterfactual questions (e.g., “What if you took that route instead of the current route?”) and descriptive questions (e.g., “Where will we be in ten minutes?”).

At the highest level, we see that the overall repertoire of operations of autonomous vehicles are strongly integrated with each other. Several sensors firstly power the vision ability of the vehicle. Then some set of sensors helps vehicle understand its precise position on the road. And as next steps, vehicle plans its motion along the trajectory and its controller adjusts the motion from the initial point to the terminal point. The general structure of AVs operations are depicted in Figure 4. It is also important to note that automotive industries have different approaches to selecting sensor technology and AI techniques to collect, fuse and process driving-related information. For example, Audi A8 2018 uses radar sensors, a front camera, ultrasonic sensors, and a laser scanner to perceive its surroundings and drive accordingly [127]. Initially, this model also used a system called “Traffic Jam Pilot,” a built-in AI control button that automatically takes control of the vehicle from a driver if the vehicle’s speed is lower than 60 km/h, in slow-moving traffic. However, the company later announced that they would not supply the new model automobiles, such as the high-end A8 Sedan, with such a control mechanism [128]. Tesla, another well-known autonomous vehicle manufacturer, uses 8 external cameras, 12 ultrasonic sensors, and a powerful onboard computer to drive their cars in a semi-autonomous fashion [129]. Their chief AI scientist recently clarified that Tesla’s ADAS, called “Autopilot,” solely uses video input to perceive the driving environment and avoids using LIDAR sensors. The rationale behind such a decision is that before creating an HD map, one has to pre-map the surroundings with LIDAR. It becomes computationally extremely convoluted

and unfavorable to produce a precise map of all locations a vehicle travels within, as explained by the vision expert at a recent computer vision and pattern recognition (CVPR) 2021 workshop on autonomous driving [130]. On the contrary, another autonomous vehicle company, Google’s Waymo, uses LIDAR sensors in addition to high-resolution cameras and radar sensors to plan a safe journey [131]. These three examples show that the design of autonomous driving systems in terms of sensor fusion and relevant AI approaches varies significantly. This is presumably a consequence of resource choices and the relative efficiency of such choices in the variety of autonomous cars regarding their target local infrastructure, road ecosystem, and road culture between people and vehicles. The prioritization of the selection of sensors is beyond the scope of this study. Still, instead, we propose XAI approaches for autonomous driving, considering safety, explainability, and regulatory compliance components in the learned controller software architecture.

VIII. AN XAI FRAMEWORK: INTEGRATING AUTONOMOUS CONTROL, EXPLANATION, AND REGULATORY COMPLIANCE

There is already a growing number of survey articles on explainable AI for autonomous systems, such as [43], [132], [133]. Our intent is not to present yet another summary survey but instead to sketch a framework in which methods for developing XAI, end-to-end autonomous systems, and regulatory compliance are combined to inform general processes of regulatory principles. Each of the three components has a role in our framework. We have already proposed 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

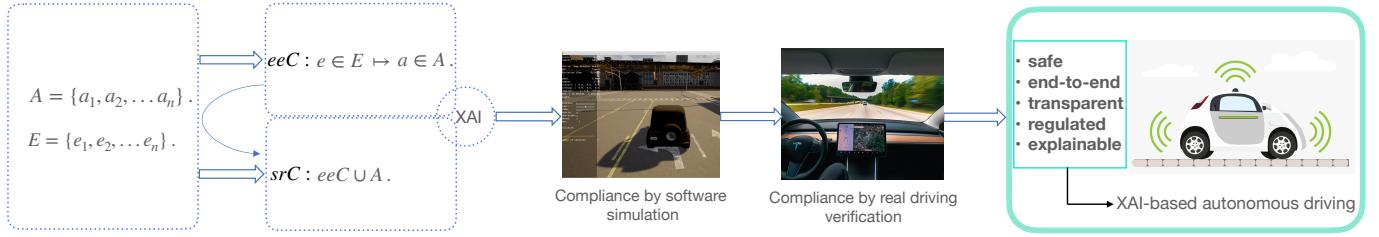


Fig. 5: A conceptual framework of XAI-based autonomous driving. An image segment in the simulation step taken from [134].

autonomous driving:

1. An end-to-end autonomous control system component:

Given all possible instances of environment,

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

and a compendium of actions

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

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

$$C : E \mapsto A.$$

This mapping intends to ensure that a controller that maps the environment to a relevant action function 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. 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 [52] and [59].

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., [135]). 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 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., [136]). The general idea is that an algorithm, for example an end-to-end autonomous control model or our eeC , can be proven to be correct for all appropriate actions. In our case, the correctness would be confirmed if all the mappings from 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 or secure 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. Bearing the concepts of the proposed framework in mind, we define XAI for autonomous driving as follows:

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

A simple graphical illustration of the aforementioned framework can be seen at Figure 5.

The state-of-the-art end-to-end learning examples [137], [138] and more recent work on empirical successes of formal verification of safe driving [139], [140] show that autonomous driving has achieved a major landmark in safety, one of the most vital driving features. However, the concept of explainability remains unsettled in the literature, and should be the next indispensable direction to fully comply with the driving regulations. What is the best way to achieve the explainability of an AI system in autonomous driving? Should the research priorities be solely devoted to post-hoc explanations, or to constructing computational models providing outcomes and direct explanations on top of such computations? One of the seminal studies on this topic is Rudin’s [141] 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. We show with examples that both concurrent and post-hoc explanations are favorable for autonomous driving. From this perspective, we present an outlook on AI methods which provide a framework for a certified software system that fulfills the requirements of transportation regulations.

IX. MIND THE GAP: FUTURE PERSPECTIVE OF XAI IN AUTONOMOUS DRIVING

In this section we propose a set of approaches, which we call a *field guide*, to guide the pursuit of the goals of XAI for autonomous driving within the principle of the presented framework. We note the recent increase in the role of XAI across many applications in both industrial and academic settings (e.g., [142], [143], [144]). In this context, we shed light on the AI approaches that can explain a vehicle’s perception system and which provide understandable real-time state-action mappings.

A. Explainable Vision

1) *Explainable vision through post-hoc explanations:* Explainable vision-directed actions for autonomous vehicles are



Fig. 6: 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 [145].

based on how high-level features are used to detect and identify objects. CNNs, as deep learning architectures, are “black-box” models; there is a need to develop explainable CNN (XCNN) architectures to comply with the requirements of the proposed framework. The studies in Section VII.A, such as DeepLift, CAM, and its variations, are promising attempts to interpret CNN predictions. It is worthwhile to further test XCNN architectures on self-driving vehicles. XCNNs may be a major step towards attaining explainable and trusted vision models in autonomous machines, by providing intelligible and trustable explanations which are coupled with accurate environmental perception. 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?” [146]. 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 for post-accident investigation under regulatory compliance. The second example is the hacking of Tesla’s Autopilot by the McAfee Advanced Threat Research team [145]. 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 6). 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 taken could be helpful to produce reliable post-hoc explanations for critical traffic scenarios.

2) *Explainable vision through computation-based concurrent explanations:* As explained in Challenge 3 in [141], developing interpretable models that concurrently provide explanations is a promising approach towards achieving a transparent vision system, particularly in object classification tasks. Initial specifications on this perspective use a concept of the *prototype layer* as an addition to a deep neural network: parts of a visual object are decomposed into pieces, and the prototype layer picks out some representative parts of the object during training. When given a new image during testing, the network tries to discover the similarities between those parts of the new image and the ones learned as prototypes in training. In this way, the deep network accumulates evidence from the prototypes and classifies the objects accordingly. Empirical studies of this perspective on classifying handwritten digits, cars, Fashion-MNIST dataset [147], and bird species and cars [148] 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 a further post-hoc explanation. In the context of autonomous driving, generating interpretable model-based concurrent explanations can contribute significantly to accident prevention. We support this proposition with a specific example. Assume that a self-driving vehicle has an in-vehicle person (i.e., a backup driver or a passenger). The vehicle provides a control (i.e., stop) button for emergency use. The in-vehicle interface shows there is *no* human ahead crossing the road and continues driving; but there *is* an on-road human ahead (i.e., a vision system malfunction). Then 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 concurrent explanations has potential use in autonomous driving and provides an opportunity for safe navigation of a vehicle.

B. Explainable State-Action Mapping Using 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 Markov Decision Processes (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 [149]. Hence, driving policies obtained under this setting can not be controlled at testing time [150]. Model-free

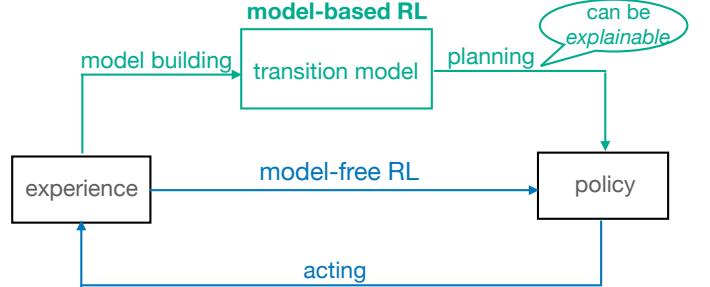


Fig. 7: Model-based RL vs. model-free RL from explainability perspective.

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 [151]. Such exploration lacks the explainability of a learned policy. A model-free RL agent will casually 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 addressed in model-based RL: An advantage of model-based RL is that an agent learns the model of the environment first and adjusts its learning policy according to the dynamics of the environment [152], [153], [154]. Such a 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 [155], [156], [157]. The Dyna and its variation, the linear-Dyna architecture, concurrently learn a world model while learning the optimal policy through interacting with the world. Dyna’s planning process creates a predicted future trajectory from an initially provided imaginary state. Based on such 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 an explanatory basis. 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,” and “go straight” actions. As prior knowledge, these four predictions can be deemed 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 [158], [159], [160], [161]. Moreover, such knowledge is also a reliable basis for explanations of the po-

tential actions taken by an autonomous system. In this context, general value functions (GVFs) are preliminary techniques for representing predictive knowledge. By definition, GVF_s aim to capture long-term predictive summaries about actual observations of an RL agent [161], [162]. For instance, an RL agent in autonomous 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 proven GVF_s’s representation value in learning, and their computational accuracy as a proof of concept has been established [161]. Moreover, recent applications of predictive knowledge, formalized as GVF_s, on perception problems [163], and learning policies in the real-world autonomous driving [164] 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 learned model 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?” or “Do the pedestrians ahead intend to cross the road?” or “Will the car ahead accelerate?” 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 needs to be question-driven. 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 *temporal actions* taken. A suitable RL approach corresponding to such an architecture is the concept of *options* [165]. Options are generalization of actions in which RL agent has a policy for taking an action with the *terminal* state. The recently proposed *option-critic architecture* is motivated by the concept of options [166]. 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 risks. 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 [167] have shown that considering no risks but only reward as a metric as in traditional RL is not always the perfect decision for an autonomous system, and 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* with respect to appropriate actions can help achieve timely, intuitive, informative, and trustworthy explanations of intelligent vehicles in critical traffic circumstances.

X. CONCLUSIONS

This study presented a systematic overview and 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 transparency in real-time decisions. In this regard, we first identified the current gaps, issues, and concerns from legal and societal perspectives. We also showed how autonomous vehicles interact with the driving environment and make real-time decisions with the help of sensor technologies. Considering the existing issues, we proposed an end-to-end framework indicating essential factors and constituents leading to confirmed safe and transparent autonomous driving. Finally, we advocated for a set of AI approaches that can help achieve regulatory compliant autonomous driving within the principles of the proposed framework. Our work can be considered as a *cause-effect-solution* perspective. We elaborate on *cause* through current gaps, concerns, and 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 a field guide that incorporates XAI principles for autonomous driving. If the proposed guidelines are properly implemented, we can move a step closer to safer, transparent, publicly approved, and environmentally-friendly intelligent automobiles in the near future.

ACKNOWLEDGMENT

We acknowledge support from the Alberta Machine Intelligence Institute (Amii), from the Computing Science Department of the University of Alberta, and the Natural Sciences and Engineering Research Council of Canada (NSERC). Shahin Atakishiyev also acknowledges support from the Ministry of 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. Bucsuhaý, E. Matuchová, R. Zúvala, P. Moravcová, M. Kostíková, 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. Katrakazas, M. Quddus, 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. Skrčkij, E. Šabanović, 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 *et al.*, "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. Yurtsever, 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. Krpalcova, 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] 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.
- [18] J. Shuttleworth, "SAE standard news: J3016 automated-driving graphic update," <https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic>, 2019, Accessed online on August 16, 2021.
- [19] 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.
- [20] J. Fleetwood, "Public health, ethics, and autonomous vehicles," *American Journal of Public Health*, vol. 107, no. 4, pp. 532–537, 2017.
- [21] P. Foot, "The problem of abortion and the doctrine of the double effect," *Oxford review*, vol. 5, 1967.
- [22] J. Jarvis Thomson, "The trolley problem," *Yale Law Journal*, vol. 94, no. 6, p. 5, 1985.
- [23] P. E. Ross, "The Audi A8: the world's first production car to achieve level 3 autonomy," *IEEE Spectrum*, vol. 1, 2017.
- [24] E. Ackerman, "What full autonomy means for the waymo driver," *IEEE Spectrum*, vol. 4, 2021.
- [25] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.
- [26] B. Lundgren, "Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles," *AI & SOCIETY*, pp. 1–11, 2020.
- [27] J. Harris, "The immoral machine," *Cambridge Quarterly of Healthcare Ethics*, vol. 29, no. 1, pp. 71–79, 2020.
- [28] S. Sohrabi, H. Khreis, and D. Lord, "Impacts of autonomous vehicles on public health: a conceptual model and policy recommendations," *Sustainable Cities and Society*, vol. 63, p. 102457, 2020.
- [29] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, "Ethical issues in focus by the autonomous vehicles industry," *Transport Reviews*, pp. 1–22, 2021.
- [30] 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.
- [31] L. Collingwood, "Privacy implications and liability issues of autonomous vehicles," *Information & Communications Technology Law*, vol. 26, no. 1, pp. 32–45, 2017.
- [32] A. Taeihagh 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.
- [33] N. Saqib, M. M. Yousuf, and M. Rashid, "Design and implementation issues in autonomous vehicles-a comparative review," in *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. IEEE, 2021, pp. 157–162.
- [34] T. Imai, "Legal regulation of autonomous driving technology: Current conditions and issues in Japan," *IATSS Research*, vol. 43, no. 4, pp. 263–267, 2019.
- [35] 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.
- [36] 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.
- [37] I. Yaqoob, L. 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.
- [38] S. Pettigrew, J. D. Nelson, and R. Norman, "Autonomous vehicles and cycling: Policy implications and management issues," *Transportation Research Interdisciplinary Perspectives*, vol. 7, p. 100188, 2020.
- [39] M. Koschuch, W. Sebron, Z. Szalay, Á. Török, H. Tschiürtz, and I. Wahl, "Safety & security in the context of autonomous driving," in *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*. IEEE, 2019, pp. 1–7.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [44] P. Langley, "Varieties of explainable agency," in *ICAPS Workshop on Explainable AI Planning (XAIP)*, 2019.
- [45] 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.
- [46] D. Omeiza, H. Webb, M. Jirocka, and L. Kunze, "Explanations in autonomous driving: A survey," *arXiv preprint arXiv:2103.05154*, 2021.
- [47] 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.
- [48] T. R. Roth-Berghofer, "Explanations and case-based reasoning: Foundational issues," in *European Conference on Case-Based Reasoning*. Springer, 2004, pp. 389–403.
- [49] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, 2009, pp. 195–204.
- [50] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of*

- the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.
- [51] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, “Designing theory-driven user-centric explainable ai,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.
- [52] GDPR, “Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016,” *Official Journal of the European Union*, 2016.
- [53] 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 29-September-2021.
- [54] 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.
- [55] National Highway Traffic Safety Administration, *Federal automated vehicles policy: Accelerating the next revolution in roadway safety*. US Department of Transportation, 2016.
- [56] 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.
- [57] NTSB, “Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016,” 2017.
- [58] NTSB Board, “Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Mountain View, California, March 23, 2018,” 2020.
- [59] Transport Canada, “Guidelines for testing automated driving systems in Canada,” *The Ministry of Transportation of Canada*, 2021.
- [60] Bundesanzeiger Verlag Board, “Act amending the road traffic act and the compulsory insurance act (autonomous driving act),” 2021.
- [61] Department for Transport team, “Safe use of automated lane keeping system (alks) summary of responses and next steps,” 2021.
- [62] Austroads, *Guidelines for trials of automated vehicles in Australia*. National Transport Commission, 2017.
- [63] Advanced information and telecommunications network society, *Outline of systematic preparations related to autonomous driving*. The Government of Japan, 2017.
- [64] ISO 21448 committee. Road vehicles — Safety of the intended functionality. (Accessed on October 8, 2021). [Online]. Available: <https://www.iso.org/standard/70939.html>
- [65] ISO 26262 committee. Road vehicles — Functional safety. (Accessed on October 8, 2021). [Online]. Available: <https://www.iso.org/standard/68383.html>
- [66] Technical committees. ISO/TC 204 Intelligent transport systems. (Accessed on October 8, 2021). [Online]. Available: <https://www.iso.org/committee/54706.html>
- [67] CEN/TC 278 Committee. Intelligent Transport Systems. (Accessed on October 8, 2021). [Online]. Available: <https://www.itsstandards.eu/>
- [68] 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>
- [69] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, 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.
- [70] M. N. Ahangar, Q. Z. 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.
- [71] F. Rosique, P. J. Navarro, C. Fernández, and A. Padilla, “A systematic review of perception system and simulators for autonomous vehicles research,” *Sensors*, vol. 19, no. 3, p. 648, 2019.
- [72] Z. Wang, Y. Wu, and Q. Niu, “Multi-sensor fusion in automated driving: A survey,” *IEEE Access*, vol. 8, pp. 2847–2868, 2019.
- [73] M.-g. Cho, “A study on the obstacle recognition for autonomous driving RC car using lidar and thermal infrared camera,” in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2019, pp. 544–546.
- [74] J. Kim, A. Rohrbach, T. Darrell, J. Cann, and Z. Akata, “Textual explanations for self-driving vehicles,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 563–578.
- [75] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, “Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7699–7707.
- [76] T. You and B. Han, “Traffic accident benchmark for causality recognition,” in *European Conference on Computer Vision*. Springer, 2020, pp. 540–556.
- [77] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, “Explainable object-induced action decision for autonomous vehicles,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9523–9532.
- [78] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [79] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, and B. N. Dugger, “Interpretable classification of alzheimer’s disease pathologies with a convolutional neural network pipeline,” *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.
- [80] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international Conference on Computer Vision*, 2017, pp. 618–626.
- [81] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.
- [82] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, “Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models,” *arXiv preprint arXiv:1908.01224*, 2019.
- [83] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [84] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, “Interpreting the predictions of complex ml models by layer-wise relevance propagation,” *arXiv preprint arXiv:1611.08191*, 2016.
- [85] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, “Unmasking clever hans predictors and assessing what machines really learn,” *Nature Communications*, vol. 10, no. 1, pp. 1–8, 2019.
- [86] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, “Visualbackprop: Efficient visualization of cnns for autonomous driving,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4701–4708.
- [87] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.
- [88] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 3145–3153.
- [89] H. K. B. Babiker and R. Goebel, “An Introduction to Deep Visual Explanation,” *31st Neural Information Processing Systems Conference (NIPS), Long Beach, CA, USA*, 2017.
- [90] H. Babiker and R. Goebel, “Using KL-divergence to focus Deep Visual Explanation,” *31st Neural Information Processing Systems Conference (NIPS), Interpretive ML Symposium. Long Beach, CA, USA*, 2017.
- [91] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, “Explainability of vision-based autonomous driving systems: Review and challenges,” *arXiv preprint arXiv:2101.05307*, 2021.
- [92] 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.
- [93] T. Litman, “Autonomous vehicle implementation predictions: Implications for transport planning,” 2020.
- [94] Rebecca Paredes. Best Apps For Green Travel And Slashing Fuel Consumption. (Accessed on October 28, 2021). [Online]. Available: <https://greenfuture.io/transportation/driving-route-planner-travel/>
- [95] M. Á. 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.
- [96] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, “A survey of motion planning and control techniques for self-driving urban vehicles,” *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016.
- [97] E. W. Dijkstra *et al.*, “A note on two problems in connexion with graphs,” *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

- [98] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [99] A. V. Goldberg and C. Harrelson, "Computing the shortest path: A search meets graph theory." in *SODA*, vol. 5. Citeseer, 2005, pp. 156–165.
- [100] R. Geisberger, P. Sanders, D. Schultes, and C. Vetter, "Exact routing in large road networks using contraction hierarchies," *Transportation Science*, vol. 46, no. 3, pp. 388–404, 2012.
- [101] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 187–210, 2018.
- [102] M. Buehler, K. Iagnemma, and S. Singh, *The DARPA urban challenge: autonomous vehicles in city traffic*. Springer, 2009, vol. 56.
- [103] A. Bacha, C. Baumran, R. Faruque, M. Fleming, C. Terwelp, C. Reinholz, D. Hong, A. Wicks, T. Alberi, D. Anderson *et al.*, "Odin: Team victortango's entry in the darpa urban challenge," *Journal of Field Robotics*, vol. 25, no. 8, pp. 467–492, 2008.
- [104] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Etinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhnke *et al.*, "Junior: The stanford entry in the urban challenge," *Journal of Field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.
- [105] J. Bohren, T. Foote, J. Keller, A. Kushleyev, D. Lee, A. Stewart, P. Vernaza, J. Derenick, J. Spletzer, and B. Satterfield, "Little ben: The ben franklin racing team's entry in the 2007 darpa urban challenge," *Journal of Field Robotics*, vol. 25, no. 9, pp. 598–614, 2008.
- [106] C. R. Baker and J. M. Dolan, "Traffic interaction in the urban challenge: Putting boss on its best behavior," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 1752–1758.
- [107] T. Wongpiromsarn, U. Topcu, N. Ozay, H. Xu, and R. M. Murray, "Tulip: a software toolbox for receding horizon temporal logic planning," in *Proceedings of the 14th International Conference on Hybrid Systems: Computation and Control*, 2011, pp. 313–314.
- [108] M. Bouton, J. Karlsson, A. Nakhaei, K. Fujimura, M. J. Kochenderfer, and J. Tumova, "Reinforcement learning with probabilistic guarantees for autonomous driving," *arXiv preprint arXiv:1904.07189*, 2019.
- [109] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 87–96, 2020.
- [110] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9704–9710.
- [111] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [112] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.
- [113] A. Pal, S. Mondal, and H. I. Christensen, "" looking at the right stuff"-guided semantic-gaze for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 883–11 892.
- [114] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2019.
- [115] J. Bidot, S. Biundo, T. Heinroth, W. Minker, F. Nothdurft, and B. Schattenberg, "Verbal plan explanations for hybrid planning." in *MKWI*. Citeseer, 2010, pp. 2309–2320.
- [116] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," *arXiv preprint arXiv:1709.10256*, 2017.
- [117] R. Borgo, M. Cashmore, and D. Magazzeni, "Towards providing explanations for ai planner decisions," *arXiv preprint arXiv:1810.06338*, 2018.
- [118] R. Korpan and S. L. Epstein, "Toward natural explanations for a robot's navigation plans," *Notes from the Explainable Robotic Systems Workshop, Human-Robot Interaction*, 2018.
- [119] B. Krarup, M. Cashmore, D. Magazzeni, and T. Miller, "Model-based contrastive explanations for explainable planning," 2019.
- [120] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable automated planning & decision making." in *IJCAI*, 2020, pp. 4803–4811.
- [121] C. Liu, A. Carvalho, G. Schildbach, and J. K. Hedrick, "Stochastic predictive control for lane keeping assistance systems using a linear time-varying model," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 3355–3360.
- [122] P. Falcone, F. Borrelli, J. Asgari, H. E. Tseng, and D. Hrovat, "A model predictive control approach for combined braking and steering in autonomous vehicles," in *2007 Mediterranean Conference on Control & Automation*. IEEE, 2007, pp. 1–6.
- [123] F. Borrelli, A. Bemporad, M. Fodor, and D. Hrovat, "An mpc/hybrid system approach to traction control," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 3, pp. 541–552, 2006.
- [124] A. Ziebinski, R. Cupek, D. Grzecha, and L. Chruszczyk, "Review of advanced driver assistance systems (adas)," in *AIP Conference Proceedings*, vol. 1906, no. 1. AIP Publishing LLC, p. 120002.
- [125] 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-a7292621135/>
- [126] Mobileye's Blog. Mobileye SuperVision™ for Hands-free ADAS. (Accessed on November 1, 2021). [Online]. Available: <https://www.mobileye.com/super-vision/>
- [127] Audi Media Center. The new Audi A8 – conditional automated at level 3 . (Accessed on November 7, 2021).
- [128] Autonomous News Europe Blog. Audi quits bid to give A8 Level 3 autonomy. (Accessed on November 7, 2021).
- [129] Tesla's Blog. Future of Driving. (Accessed on November 7, 2021). [Online]. Available: https://www.tesla.com/en_CA/autopilot
- [130] CVPR 2021 news. CVPR 2021 Workshop on Autonomous Driving. (Accessed on November 7, 2021). [Online]. Available: <http://cvpr2021.wad.vision/>
- [131] Waymo's Blog. Waymo Driver. (Accessed on November 7, 2021). [Online]. Available: <https://waymo.com/waymo-driver/>
- [132] M. Winikoff, "Towards trusting autonomous systems," in *International Workshop on Engineering Multi-Agent Systems*. Springer, 2017, pp. 3–20.
- [133] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *arXiv preprint arXiv:2006.00093*, 2020.
- [134] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*. PMLR, 2017, pp. 1–16.
- [135] 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.
- [136] S. Hantler and J. King, "An introduction to proving the correctness of programs," *ACM Comput. Surv.*, vol. 8, pp. 331–353, 09 1976.
- [137] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.
- [138] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7153–7162.
- [139] C. Pek, S. Manzinger, M. Koschi, and M. Althoff, "Using online verification to prevent autonomous vehicles from causing accidents," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 518–528, 2020.
- [140] 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.
- [141] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [142] R. Voster and R. Slagter. (2017) Autonomous compliance: standing on the shoulders of Regtech! [Online]. Available: <https://www.compact.nl/articles/autonomous-compliance/>
- [143] Global Mining Guidelines Group. Guideline for the Implementation of Autonomous Systems in Mining. [Online]. Available: https://gmgroup.org/wp-content/uploads/2019/06/20181008_Implementation_of_Autonomous_Systems-GMG-AM-v01-r01.pdf
- [144] M. Fisher, V. Mascardi, K. Y. Rozier, B.-H. Schlingloff, M. Winikoff, and N. Yorke-Smith, "Towards a framework for certification of reliable autonomous systems," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 1, pp. 1–65, 2021.
- [145] S. Povolny and S. Trivedi, "Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles," *McAfee Advanced Threat Research*, 2020.

- [146] ITU Team, “The Molly Problem,” *AI for autonomous and assisted driving*, (Accessed on November 23, 2021). [Online]. Available: <https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx>
- [147] O. Li, H. Liu, C. Chen, and C. Rudin, “Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [148] 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.
- [149] Z. Xu, J. Chen, and M. Tomizuka, “Guided policy search model-based reinforcement learning for urban autonomous driving,” *arXiv preprint arXiv:2005.03076*, 2020.
- [150] 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.
- [151] R. S. Sutton and A. G. Barto, *Reinforcement learning: An Introduction*. MIT Press, 2018.
- [152] H. Yao and C. Szepesvári, “Approximate policy iteration with linear action models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012.
- [153] H. Yao, C. Szepesvári, B. A. Pires, and X. Zhang, “Pseudo-mdps and factored linear action models,” in *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, 2014, pp. 1–9.
- [154] T. M. Moerland, J. Broekens, and C. M. Jonker, “Model-based reinforcement learning: A survey,” *arXiv preprint arXiv:2006.16712*, 2020.
- [155] R. S. Sutton, “Dyna, an integrated architecture for learning, planning, and reacting,” *ACM Sigart Bulletin*, vol. 2, no. 4, pp. 160–163, 1991.
- [156] R. S. Sutton, C. Szepesvári, A. Geramifard, and M. Bowling, “Dyna-style planning with linear function approximation and prioritized sweeping,” in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, ser. UAI’08. Arlington, Virginia, USA: AUAI Press, 2008, p. 528–536.
- [157] H. Yao, S. Bhatnagar, D. Diao, R. S. Sutton, and C. Szepesvári, “Multi-step dyna planning for policy evaluation and control,” in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22. Curran Associates, Inc., 2009. [Online]. Available: <https://proceedings.neurips.cc/paper/2009/file/c52f1bd66cc19d05628bd8bf27af3ad6-Paper.pdf>
- [158] G. L. Drescher, *Made-up minds: a constructivist approach to artificial intelligence*. MIT Press, 1991.
- [159] R. S. Sutton and B. Tanner, “Temporal-difference networks,” *Advances in Neural Information Processing Systems*, vol. 17, 2004.
- [160] 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 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 2011, pp. 761–768.
- [161] A. White, “Developing a predictive approach to knowledge,” *PhD thesis, University of Alberta*, 2015.
- [162] G. Comanici, D. Precup, A. Barreto, D. K. Toyama, E. Aygün, P. Hamel, S. Vezhnevets, S. Hou, and S. Mourad, “Knowledge representation for reinforcement learning using general value functions,” 2018.
- [163] D. Graves, K. Rezaee, and S. Scheideman, “Perception as prediction using general value functions in autonomous driving applications,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1202–1209.
- [164] D. Graves, N. M. Nguyen, K. Hassanzadeh, and J. Jin, “Learning predictive representations in autonomous driving to improve deep reinforcement learning,” *arXiv preprint arXiv:2006.15110*, 2020.
- [165] R. S. Sutton, D. Precup, and S. Singh, “Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning,” *Artificial Intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.
- [166] P.-L. Bacon, J. Harb, and D. Precup, “The option-critic architecture,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [167] S. Zhang and H. Yao, “QUOTA: The quantile option architecture for reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5797–5804.



Shahin Atakishiyev is from Mirzabayli 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.