IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 June 2024, accepted 14 July 2024, date of publication 19 July 2024, date of current version 30 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3431437

RESEARCH ARTICLE

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

SHAHIN ATAKISHIYEV[1], (Graduate Student Member, IEEE), MOHAMMAD SALAMEH[2], HENGSHUAI YAO[3], AND RANDY GOEBEL[1]

[1]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada
[2]Huawei Technologies Canada Company Ltd., Edmonton, AB T6G 2C8, Canada
[3]Sony AI, Edmonton, AB, Canada

Corresponding author: Shahin Atakishiyev (shahin.atakishiyev@ualberta.ca)

This work was supported in part by Alberta Machine Intelligence Institute (Amii); in part by the Computing Science Department, University of Alberta; and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The work of Shahin Atakishiyev was supported by the Ministry of Science and Education of the Republic of Azerbaijan.

ABSTRACT Autonomous driving has achieved significant milestones in research and development over the last two decades. There is increasing interest in the field as the deployment of autonomous vehicles (AVs) promises safer and more ecologically friendly transportation systems. With the rapid progress in computationally powerful artificial intelligence (AI) techniques, AVs can sense their environment with high precision, make safe real-time decisions, and operate reliably without human intervention. However, intelligent decision-making in such vehicles 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, AVs must also explain their AI-guided decision-making process in order to be regulatory-compliant across many jurisdictions. Our study sheds comprehensive light on the development of explainable artificial intelligence (XAI) approaches for AVs. In particular, we make the following contributions. First, we provide a thorough overview of the state-of-the-art and emerging approaches for XAI-based autonomous driving. We then propose a conceptual framework considering the essential elements for explainable end-to-end autonomous driving. Finally, we present XAI-based prospective directions and emerging paradigms for future directions that hold promise for enhancing transparency, trustworthiness, and societal acceptance of AVs.

INDEX TERMS Autonomous driving, explainable artificial intelligence, intelligent transportation systems, regulatory compliance, safety.

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]. Such a lack of rule obedience and improper road culture have, therefore, motivated officials, manufacturers, and legislators to make substantial improvements in transportation systems. In this sense, there are growing research and development attempts to enhance safety and automation capability of AVs with the goal of preventing traffic accidents, and creating a better road infrastructure. Intel's report on the projected benefits of AVs estimates that deployment of this technology on roads will result in a reduction of 250 million hours of users' commuting time per year and save more than half a million lives from 2035 to 2045, just in the USA [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Yunlong Cai.

While the potential impact and benefits of AVs in everyday life are promising, there is a major societal concern about functional safety of such vehicles. This issue, as a major drawback, originates mainly from reports of recent traffic accidents with the presence of AVs, primarily owing to their "black-box" decision-making [3], [4], [5], [6]. As AI approaches provide the foundation for real-time driving actions, there is an inherent need and expectation from consumers, general society, and regulatory bodies that AI-based action decisions of AVs should be explainable to build confidence in these vehicles [3], [7], [8], [9], [10] (e.g., Figure 1).

In this survey, we present a comprehensive overview of state-of-the-art investigations on the explainability of autonomous driving. Through extensive analyses, we first show the background information on the need for the emergence of explanations for AVs. Furthermore, we fill the gap in the current literature by providing a structured and comprehensive review of state-of-the-art and emerging XAI approaches for autonomous driving and present a road map for future directions. More specifically, we discuss the following research questions in depth:

1) Why is there a need for XAI in AV technology?
2) What are the current trends and emerging AI technologies for explainable autonomous driving?
3) What are promising future XAI directions toward trustworthy, responsible, regulatory-compliant, and publicly acceptable AVs?

With these questions in mind, our paper makes the following contributions:

- We describe cross-disciplinary perspectives and requirements necessitating explainability in autonomous driving;
- We provide a survey of state-of-the-art XAI-based investigations for autonomous driving;
- We present a conceptual framework for explainable end-to-end autonomous driving;
- We propose future research directions on promising XAI approaches for autonomous driving.

The rest of the article consists of six sections. Section II provides background information and the factors triggering the need for the emergence of XAI in autonomous driving. Section III covers the concept of explainability for AVs by analyzing 1) cross-disciplinary perspectives necessitating explanations, 2) the role of various types of explanations for diverse explanation recipients, and 3) construction methodologies for explanations. Section IV provides a comprehensive survey of studies on XAI-based autonomous driving. Motivated by current limitations and trends delineated in these works, Section V presents a general design framework for explainable autonomous driving and shows key components of such a framework. Finally, Section VI outlines potential challenges and a road map toward safety and explainability of next-generation AVs, as future directions, and Section VII concludes the paper with an overall summary.

FIGURE 1. A canonical example of explainable AI in autonomous driving: An autonomous vehicle provides a live natural language explanation of its real-time decision to bystanders at a crosswalk. The image has been adapted and modified from the original source: [15].

II. BACKGROUND

This section describes background information on AVs with autonomous driving at a glance, fundamental issues, and regulations and standards within AVs.

A. AUTONOMOUS DRIVING AT A GLANCE

AVs, also referred to as self-driving vehicles, are intelligent vehicles equipped with advanced sensors, cameras, RADAR, LIDAR, GPS, and sophisticated learning algorithms that enable them to navigate and operate without human intervention [11]. To discern, identify, and distinguish the objects in their operational surroundings, these vehicles fuse information from a variety of sensors that help make real-time driving decisions [12], [13]. The history of contemporary AVs goes back to 1988, when ALVINN (Autonomous Land Vehicle In a Neural Network), the first neural network-powered self-driving vehicle taking camera images with a laser range finder, was able to produce control commands for the road-following task [14]. Current AVs deployed on road networks have different levels of automation based on their in-vehicle technologies and intelligent capabilities. SAE International (previously known as the Society of Automotive Engineers) has defined six levels of autonomous driving [16]: Level 0 - No automation (a human driver is responsible for all critical driving tasks); Level 1 - Driving assistance (a vehicle has automated driving support such as acceleration/braking or steering, but the driver is responsible for all other possible driving operations); Level 2 - Partial automation (Advanced Driving Assistance Systems (ADAS) operations such as steering and acceleration/braking are available in this level); Level 3 - Conditional automation (a vehicle has more advanced features such as object/obstacle detection and can carry out the majority of driving operations); Level 4 - High automation (a vehicle can fulfill all possible driving operations in a geofenced area); and Level 5 - Full automation (a vehicle can perform all driving operations in any likely scenario, and no human intervention is required).

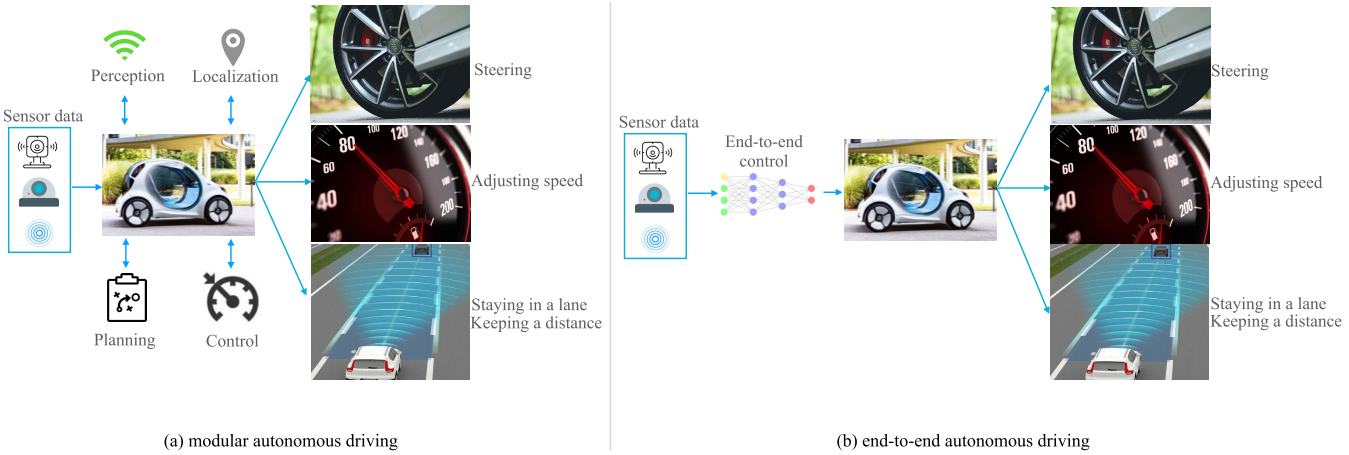

FIGURE 2. Modular vs. end-to-end autonomous driving. In the modular pipeline, the described operations are carried out subsequently to produce control commands, while end-to-end driving directly inputs raw sensor data and produces control commands as a single, unified task.

There are two main approaches to building autonomous driving systems in terms of their AI-based learning architecture: modular and end-to-end pipelines [5], [17]. The modular pipeline consists of four primary and interconnected modules categorized as *perception*, *localization*, *planning*, and *control* (Figure 2, a). The modular pipeline leverages various sensor suites and algorithms for each module. While being comprised of standalone components makes the modular system more explainable and debuggable, such an architecture propagates errors to the next component, and thus, the overall pipeline error becomes cumulative [17], [18], [19].

In contrast to a modular pipeline, end-to-end autonomous driving has recently emerged as a paradigm shift in the design and development of AVs. End-to-end autonomous driving takes the raw sensor data as visual input and yields a control command for the vehicle (Figure 2, b) [17], [20], [21]. Particularly, recent breakthroughs in deep learning and computer vision algorithms, and the availability of rich sensor devices along with enhanced safety benefits have been the primary reasons for automotive researchers to leverage the end-to-end learning approach. The advantage of an end-to-end pipeline over its counterpart is that it directly produces driving actions by unifying perception, localization, planning, and control as one combined machine learning (ML) task. Furthermore, computational efficiency is improved via shared backbones in end-to-end learning, and in this way, potential information loss in intermediate layers is also avoided [17], [22].

B. FUNDAMENTAL ISSUES

AI approaches, which are currently predominated by deep learning algorithms, have brought considerable improvements to many essential components of autonomous driving technology, including advances in perception, object detection, and planning. As the AI-powered driving systems of vehicles advance, the number of AVs deployed to road networks has proliferated significantly in many developed European countries, the US, and Canada over the last decade [23]. 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 AVs on roads. For example, Fleetwood [24] has investigated public health and ethical issues arising from the use of autonomous driving. Their study provides an in-depth analysis of the health issues, especially with the Trolley problem examples [25] and [26] (hitting a pedestrian on an icy road or a parked car; driving and hitting five people or changing the direction of the steering wheel and hitting an individual, etc.). Some studies have directly focused on the concept of ethical crashing (i.e., if crashing is inevitable, how to crash?) and the Trolley problem mentioned above. For instance, the Moral Machine experiment [27], a well-known and hotly debated experiment, investigates a general community's preferences on applied Trolley problems (inevitable accident scenarios with binary outcomes) and states that "these preferences can contribute to developing global, socially acceptable principles for machine ethics." However, further discussion on this issue condemns this opinion and draws attention to the lack of safety principles [28], which force deeper consideration of such dilemmas [29]. 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*, which arises when an accident happens and the responsibility of either an autonomous system or a human is the cause of this accident remains unresolved. Finally, there is the question of who is responsible for compensating the injured during an accident, which precipitates the third issue: the *liability gap*. That

study also shows that the core of these issues is associated with domain complexity, system complexity, and transferring more decision-making functions from humans to autonomous systems. Further studies include the outcomes of autonomous driving technology on public health in an urban area [31] and ethical dilemmas with AVs [32]. Overall, the key findings from these studies necessitate an understanding of the causes of these issues and intrinsically give the stakeholders the right to ask ''why'' questions.

C. REGULATIONS AND STANDARDS

The issues and growing concerns caused by AI systems create the need to scrutinize the regulation of this technology. As a result, public institutions have initiated the development of regulatory frameworks to monitor the activities of data-driven systems at both a country level and internationally. The focal points of these regulations are mainly to protect the stakeholders' rights and ensure they have control over their data. For example, the General Data Protection Regulation (GDPR) of the European Union (EU) initiated guidelines to promote the ''right of an explanation'' principle for users, enacted in 2016 and taking effect in May 2018 [33]. Moreover, the EU has a specially defined strategy on Guidelines of Trustworthy AI that has seven essential requirements, namely 1) human agency, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) accountability, 6) diversity, non-discrimination, and fairness, and 7) societal and environmental well-being; these principles are all to be applied in AI-based product research and development [34]. Various organizations have recently proposed guidelines on the regulation of AVs to monitor their compliance with law enforcement. NACTO's (National Association of City Transportation Officials) statement on automated vehicles proposes nine principles to shape a policy on regulation of future generation AVs [35]. NHTSA of the US Department of Transportation has a specific federal guideline on automated vehicle policy to improve traffic safety [36]. In March 2022, NHTSA announced that automobile manufacturers would no longer have to equip fully autonomous cars with manual control elements, such as a steering wheel and braking pedals in the USA [37]. Canada [38], Germany [39], UK [40], Australia [41], and Japan [42] have also recently launched their regulations on autonomous driving technology.

While the regulations have been set out to ensure legislative norms and user demands are met, some standards provide specifications to achieve a high safety level, quality assurance, efficiency, and environmentally friendly transportation systems. The International Organization for Standardization (ISO) has adopted several standards to define the relevant issues of automated driving. Examples include the ISO 21448 [43], which specifies situation awareness standards to maintain operational safety under the ''Safety of the Intended Functionality,'' and the ISO 26262 [44] standard defined for the safety of electrical and electronic systems in production passenger vehicles, entitled as ''Road vehicles - Functional safety.'' Thorough documentation on the details of legislation, regulation, and standardization of autonomous cars can be viewed in [45].

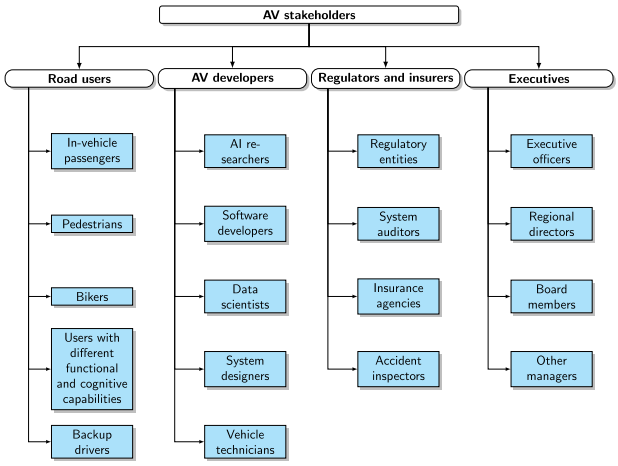

FIGURE 3. Taxonomy of the stakeholders in autonomous driving with respect to explanation conveyance.

III. EXPLANATIONS IN AUTONOMOUS DRIVING

This section describes explanations in the context of autonomous driving. It delineates the necessity of explanations, potential benefits of explanations, explanation receivers, and explanation delivery methods in AVs.

A. THE NEED FOR EXPLANATIONS IN AVS

The need for explanations in autonomous driving arises from fundamental issues, established regulations and standards covered in previous subsections, and cross-disciplinary views and opinions of society. At the highest level, the necessity of explanations for AVs can be summarized in terms of four perspectives:

Psychological perspective: Traffic accidents and safety concerns remain the main cause of the need for XAI in autonomous driving from a psychological point of view [46].

Sociotechnical perspective: The design, development, and deployment of AVs should be human-centered, reflecting the target audience's needs and taking their prior opinions and expectations into account [47], [48].

Philosophical perspective: Explaining AI decisions can provide descriptive information about the causal history of actions, particularly in critical situations [49], [50], [51].

Legal perspective: It considers all the above-mentioned factors and incorporates them into general regulatory compliance principles for AVs. A notable example is GDPR's requirements on explanation provision for end users [33].

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

B. POTENTIAL BENEFITS OF EXPLANATIONS FOR AVS

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

TABLE 1. Studies on visual explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Bojarski et al., [55], 2016	Pixel-based explanations of CNN predictions	CNN	Visual	AV developers
Kim and Canny [56], 2017	Explaining behavior of a vehicle controller using heat maps	CNN, LSTM	Visual	AV developers
Kim et al., [57], 2018	Generating textual explanations on a vehicle's control commands	CNN, S2VT, LSTM	Visual and Textual	All groups
Hofmarcher et al., [58], 2019	Visual scene understanding using semantic segmentation	Enet, SqueezeNet 1.1, ELU	Visual	AV developers
Zeng et al., [59], 2019	End-to-end interpretable neural motion planner	FaF, IntentNet	Visual	AV developers
Hu et al., [60], 2019	Interpretable multi-modal probabilistic prediction for autonomous driving	CVAE, Dynamic time warping, LSTM	Visual	AV developers
Xu et al., [61], 2020	Explaining object-induced action decisions for autonomous vehicles	Faster R-CNN	Visual	All groups
Kim et al., [62], 2020	Advisable learning for self-driving vehicles by internalizing observation-to-action rules	Mask R-CNN, LSTM	Visual and Textual	All groups
Li et al., [63], 2021	Risk object identification via causal inference	InceptionResnet-V2, Mask R-CNN, Deep SORT, RoIAlign	Textual	All groups
Casas et al., [64], 2021	End-to-end model for mapless autonomous driving	CoordConv	Visual and Textual	All groups
Kim et al., [65], 2021	Explainable and advisable model for self-driving cars	DeepLab v3, Mask R-CNN, LSTM	Textual	All groups
Wang et al., [66], 2021	Enhancing automated driving with human foresight	Gaze-based vehicle reference	Visual	Road users
Chitta et al., [67], 2021	Interpretable neural attention fields for end-to-end autonomous driving	ResNet, MLP	Visual	AV developers
Dong et al., [68], 2021	Explainable autonomous driving via an image transformer	ResNet-50, Mobilenet-v2, multi-head self-attention	Textual	All groups
Hanna et al., [69], 2021	Interpretable goal recognition in the presence of occluded factors for autonomous vehicles	Goal and Occluded Factor Inference, Monte Carlo Tree Search	Visual	AV developers
Mankodiya et al., [70], 2021	XAI for trust management in autonomous vehicles	Random Forest, Decision Tree, AdaBoost	Visual	AV developers
Madhav and Tyagi, [71], 2022	Explainable navigational intelligence for trustworthy autonomous driving	Grad-CAM, Lime	Visual	AV developers
Jing et al., [22], 2022	Interpretable action decision making for autonomous driving	Faster R-CNN	Visual and Textual	All groups
Jacob et al., [72], 2022	Region-targeted counterfactual explanations	GANs	Visual	AV developers
Zhang et al., [73], 2022	Interrelation modeling for explainable automated driving	Faster R-CNN, ResNet-50	Visual	AV developers
Kolekar et al., [74], 2022	Traffic scene understanding via U-Net and Grad-CAM	U-Net, GradCam	Visual	AV developers
Zemni et al., [75], 2023	Object-aware counterfactual explanations	BlobGAN	Visual	AV developers
Itkina and Kochenderfer [76], 2023	Trajectory prediction via interpretable self-aware neural networks	PostNet	Visual	AV developers
Feng et al., [77], 2023	Natural language explanations via semantic scene understanding	DeepLabV3	Visual and textual	All groups
Hu et al., [78], 2023	Interpretable trajectory prediction and decision-making of AVs	LaneGCNN, ResNet	Visual	AV developers
Dong et al., [8], 2023	Describing traffic scenes in natural language via attention-based transformer	CNN, LSTM, Transformer	Visual and textual	All groups
Atakishiyev et al., [79], 2023	Explaining autonomous driving actions with visual question answering	VGG-19, LSTM, DDPG	Textual	All groups
Echterhoff et al., [80], 2024	Leveraging concept bottlenecks as visual features for predicting control command and explanations of vehicle and human behavior	Longformer, GPT 3.5	Visual and textual	All groups
Feng and Sun [81], 2024	Interpreting self-driving decisions and improving safety by paying more attention to the regions that are near the ego vehicle	Multilayer Perceptron, Trajectory-guided Control Prediction	Visual	AV developers
Araluce et al., [19], 2024	Using driver attention for an end-to-end explainable decision-making from frontal driving images	ARAGAN, MobileNetV2	Visual	AV developers

1) HUMAN-CENTERED DESIGN

Getting the end users' inputs, opinions, and anticipations on the design and development of the semi or fully AVs can help with the acceptance of this technology by the general community [52].

2) TRUSTWORTHINESS

Algorithmic assurance can build trust in human-autonomous system relationships [53].

3) TRACEABILITY

Explainable intelligent driving systems can help forensic analysts and system auditors understand the entire decision-making process of an autonomous car during the journey via a post-trip analysis.

4) TRANSPARENCY AND ACCOUNTABILITY

Explanations can help achieve accountability, which can resolve the potential liability and responsibility gaps in

foreseeable post-accident investigations with the involvement of AVs as described by Burton et al. [30]. For example, Mercedes-Benz has recently taken a promising step forward and announced that the corporation will take legal responsibility for any accidents that their self-driving systems are engaged in [54]. Mercedes's declaration of legal culpability is a significant milestone toward the accountability of AV technology.

C. EXPLANATION RECIPIENTS IN AVS

The details, types, and delivery of explanations vary in accordance with users' identities, technical background knowledge in autonomous driving, and their various functional and cognitive abilities [46], [82]. For instance, a user having little technical expertise on how AVs operate may be satisfied with a simple explanation of a relevant decision/outcome. However, an autonomous systems engineer will need more informative explanations to understand the current functionalities of the car, with the motivation to appropriately ''debug'' the existing driving system as required. Therefore, the use of domain knowledge and expertise of the explainee is essential to provide pertinent, sufficiently informative, and intelligible explanations [83], [84]. Motivated by a target audience definition of [46] and [85], we can distinguish *four* groups of the stakeholders in autonomous driving, namely Group 1 - Road users, Group 2 - AV developers, Group 3 - Regulators and insurers, and Group 4 - Executive management of automobile companies. Figure 3 provides the identity of such stakeholders and their positions in the relevant classification.

D. EXPLANATION DELIVERY METHODS IN AVS

As explainees are classified based on their domain knowledge and needs, explanations and their design and evaluation techniques also vary depending on the context and knowledge of the category of explainees. In fact, explanation construction is one of the major challenges in current XAI research. Zablocki et al. [86] define four ''W'' questions in XAI-based autonomous driving: 1) Who needs explanations? 2) Why are explanations needed? 3) What kind of explanations can be generated? and 4) When should explanations be delivered? In general, explanations in AI can be distinguished based on their *derivation category* and *classification*. Some of the early practical studies have applied explanations to automated collaborative filtering systems [87] and knowledge-intensive case-based reasoning [88]. Another empirical approach has attempted to derive explanations based on some intelligibility types [89] and used ''why,'' ''why not,'' ''what if,'' and ''how to'' type explanations for causality filtering. Furthermore, Liao et al. [90] have interviewed twenty user-interface and design practitioners working in different areas of AI to understand users' explanatory requirements. By doing so, they have tried to find the gaps in the interviewers' products and developed *a question bank*: the authors represent users' needs as questions so that users may potentially ask about the outcomes produced by an AI system. Overall, the stakeholder needs-based explanation design can be viewed as one of the promising approaches for vehicular technology.

Another popular approach to producing explanations is based on using psychological tools from *formal theories*, according to the literature review of [91]. Depending on the context and addressee, both explanation derivation methods confirm their usefulness. These explanation generation approaches can find alignment in their application in autonomous driving; since autonomous driving involves people with diverse backgrounds in society, relevant XAI design needs inherent adjustments to the context problem.

Except for *informational content*, the effective communication of explanations is also a key factor for good human-machine teaming. In general, the conveyance of explanations to end users is realized through a user interface (UX) or a human-machine interface (HMI) [92]. For instance, an HMI may be an interface to alert the human driver to take over the control of a vehicle in an emergent situation. Other potential examples are text messages displayed in monitors, sound, light signals, and vibrotactile technology that explain the vehicle's intentions and bring situation awareness for people in the loop, as shown in Schneider et al.'s work [93].

IV. XAI FOR AUTONOMOUS DRIVING: A SURVEY

A. PREVIOUS SURVEYS ON XAI FOR AVS

There exist reviews on XAI for AVs, which provide valuable insights from various perspectives. These studies expose a variety of approaches, from a universal look to an algorithmic point of view. In this sense, the first noteworthy review on XAI for AVs is Omeiza et al.'s work [46]. They study the need for/role of explanations for autonomous driving and focus on legal requirements, standards, and consumer expectations for the design and development of explainable autonomous driving systems. This provides their basis for presenting a conceptual XAI framework for modular autonomous driving. In further work, Zablocki et al. [86] present a detailed overview of end-to-end vision-based autonomous driving systems and describe explainability hurdles for AVs from an ML perspective. Finally, in very recent work, Kuznietsov et al. [94] present a systematic review of XAI techniques for modular and end-to-end autonomous driving by focusing on how such techniques improve safety and user trust. In this regard, they propose the SafeX framework for modular autonomous driving by integrating user interface, safety, and explainability.

Our study, as a complement, extends the coverage of this previous work in three essential dimensions. First, all three previous surveys specifically focus on *form* and *content* of explanations; however, *time granularity* of explanations has not been investigated. As AVs are real-time decision-making systems, it is crucial to know how explanations must be delivered from the timing perspective. Furthermore, attention-based transformers, large language models, and vision-language models are now at the forefront of AI applied to AV technologies, and such approaches have not been

explored in the aforementioned surveys. Finally, a classification of XAI approaches applied to AVs from an algorithmic/methodological perspective is also a noteworthy nuance missing in these studies. Consequently, our paper extends the above-mentioned reviews by (1) analyzing the temporal sensitivity of explanations, (2) presenting a methodological taxonomy of XAI methods, and (3) providing a perspective on emerging XAI paradigms for future generation AVs.

B. STRUCTURE OF OUR PAPER

Our paper presents a comprehensive overview of XAI methods for AVs. In particular, we present a classification of approaches in terms of vision, reinforcement learning, imitation learning, feature importance, logic, user study, and the most recent paradigm shift - large language and vision-language-based explanations for AVs. By adopting insights from the recent industrial trends, emerging AI technologies, and general regulatory compliance principles, we further present a conceptual framework for explainable end-to-end autonomous driving and describe its essential elements. Finally, we present a set of promising XAI approaches by describing the missing pieces in the state of the art and present potential solutions with the goal of bridging the gaps and achieving transparency and social acceptance in the next-generation AVs.

C. VISUAL EXPLANATIONS

As deep neural networks, often in augmented forms of CNNs, power the vision ability of intelligent vehicles, understanding how CNNs capture real-time image segments that lead to the particular behavior of a vehicle is a key concept to achieving visual explanations. In this regard, explainable CNN architectures have resulted in adjustments to generate visual explanations. Zeiler and Fergus [95] use deconvolution layers to understand the internal representation of CNNs in their seminal work. Hendricks et al. [96] propose a model concentrating on distinguished properties of objects that explain the rationale for the predicted label. Zhou et al.'s [97] saliency map architecture, class activation map (CAM), highlights the discriminative part of an image to predict the label of the image. Moreover, Selvaraju et al. [98] propose an augmented version of CAM, called Grad-CAM, that highlights the derivative of CNN's prediction with respect to its input. Further examples of backpropagation-based methods include guided-backpropagation, [99], layer-wise relevance propagation [100], [101], and DeepLift [102]. Babiker and Goebel [103], [104] have also shown that heuristics-based Deep Visual Explanations (DVE) provide a grounded justification for predictions of a CNN.

Explaining autonomous driving decisions using visual techniques is also primarily motivated by these studies. Particularly, Bojarski et al.'s work [55] is the first explainable vision approach for self-driving, where the authors propose a visualization method, called VisualBackProp, showing which set of *input pixels* contributes to a prediction made by CNNs. Their experiments conducted with the Udacity self-driving car dataset on an end-to-end autonomous driving task show that the proposed technique is a useful tool for debugging predictions of CNNs.

Hofmarcher et al. [58] propose a *semantic segmentation model* implemented as a pixel-wise classification that explains underlying real-time perception of the environment. They evaluate the performance of their framework on Cityscapes [105], a benchmark dataset for understanding street scenes. The framework outperforms other popular segmentation models such as ENet and SegNet with 59.8 per-class mean intersection over union (IoU) and 84.3 per-category mean IoU. Interpretability of the model is a plus for unexpected behaviors, allowing to debug the driving system and understand the rationales for temporal decisions of a self-driving vehicle.

Kim and Canny [56] use a *causal attention* model on top of the saliency filtering that indicates which input regions actually affect the steering control. Their experiments are conducted on the driving datasets - Comma.ai [106], Udacity [107], and Hyundai Center of Excellence in Integrated Vehicle Safety Systems and Control (HCE): This model runs for nearly 16 hours to train CNNs end-to-end from images to steering angles and apply causality filtering to find out which parts of images have high influence in predictions. With this approach, the learned framework provides an interpretable visualization of a vehicle's actions. As an enhancement of this model, Kim et al. [57] provide textual explanations in their further study. They produce ''intelligible explanations'' on the decisive actions of a self-driving vehicle using an attention-based video-to-text mechanism and introduce a novel dataset, called Berkeley Deep Drive-X (eXplanation) (BDD-X), that contains annotations for textual explanations and descriptions.

Zeng et al.'s [59] architecture learns to drive an autonomous vehicle safely by following traffic rules, including interaction with road users, yielding, and traffic signals. They use raw LIDAR data and an HD map to generate interpretable representations as 3D detection of objects, anticipated future trajectories, and cost map visualizations. 3D detection instances provide descriptive information so that the model understands the operational environment. Motion forecasting, measured as L1 and L2 distances, explains whether erroneous actions are due to incorrect velocity or calculation of direction. Finally, Cost Map visualization describes the traffic scene via a top-down view. The architecture is evaluated on a large real-driving dataset consisting of 6,500 traffic scenarios with 1.4 million frames and collected across several cities in North America, measuring traffic rule violation, closeness to human trajectory, and collision. The authors also carry out an ablation study and show the impact of different overrides, input horizons, and training losses on end-to-end learning.

Xu et al. [61] propose *object-induced actions* with explanations for predictions of an autonomous car. The authors introduce a new dataset called BDD-OIA, as an extension of the BDD100K dataset [108]; this extension

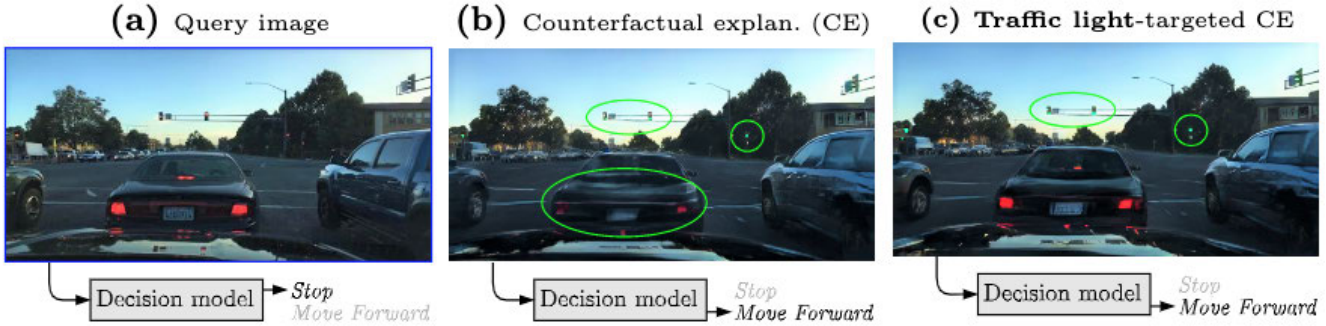

FIGURE 4. An example of a counterfactual explanation generated by STEEX. Graphics credit: [72].

is annotated with 21 explanation templates on a set of 4 actions. Their multi-task formulation for predicting actions also improves the accuracy of action selection. The CNN architecture further unifies reasoning on action-inducing objects and the context of scenes globally. The empirical results of the study on the introduced BDD-OIA dataset show that the explainability of the architecture also enhances action-inducing object recognition, resulting in better self-driving.

In two respective studies, Kim et al., [62], [65] propose an approach that leverages *human advice* to learn vehicle control (Figure 5). By sensing operational surroundings, the system is able to generate intelligible explanations on the decisive actions (For example, ''Slowing down *because* the road is wet''). The proposed architecture incorporates semantic segmentation with an attention mechanism that enriches knowledge representation. Experiments performed on the BDD-X dataset show that human advice with semantic segmentation and heat maps improves both the safety and explainability of predictive actions of a self-driving vehicle.

As a more recent vision-to-text approach, Atakishiyev et al. [79] employ the visual question answering (VQA) mechanism to explain autonomous driving actions. They train an RL agent and generate driving data showing the self-driving car's motion from its field of view. They further convert this video to image sequences, manually annotate the images with question-answer (QA) pairs, and encode questions and images with LSTM [109] and pre-trained VGG-19 [110], respectively. The experimental results on five action categories show that VQA is a straightforward, effective, and human-interpretable approach to justifying autonomous driving actions. Leveraging frontal images for interpretable decision-making has further been explored by subsequent studies [8], [19], and [80] as well.

While the mentioned studies focus on vision-based explanations of already obtained predictions of the model, there have been some recent studies paying attention to *counterfactual explanations*. In the context of automated driving, counterfactual analysis can be described with such an exemplary question: ''Given the driving scene, how can it be modified so that the vehicle keeps driving instead of stopping ?'' In other words, given the input, counterfactual

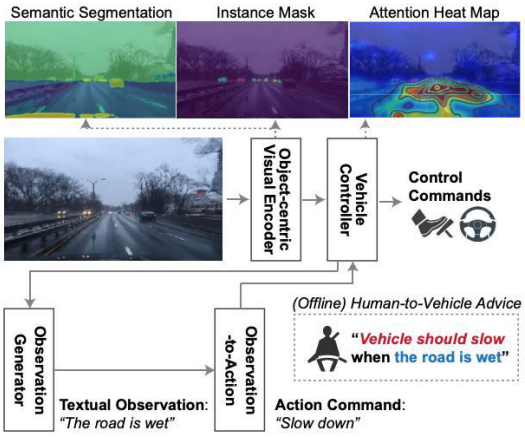

FIGURE 5. Human advice to the car for relevant action. Source: [62].

analysis intends to figure out the distinguished features in this input that cause the model to make a certain prediction by envisioning modification of those features would cause the model to make a different prediction (e.g., Figure 4). Thus, in this case, the predictions obtained by the existing model and the imagined model become contrastive. As the application of counterfactual intervention, Li et al. [63] presents an approach to find out risk objects that result in particular driving behavior. Their method, formalized as a Functional Causal Model (FCM), shows that the random elimination of some objects from the scene changes the driving decision to the contrastive prediction, such as from the ''Stop'' to ''Go'' command. In further work, Jacob et al. [72] introduce the STEEX model that uses a pre-trained generative model to produce counterfactual rationales by modifying the style of the scene while retaining the structure of the driving scene. Finally, as further enhancement of STEEX, Zemni et al. [75] propose a method called OCTET that generates object-aware counterfactual explanations without depending on the structural layout of the driving scene as backpropagation can optimize the spatial positions of the provided instances.

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

Table 1 summarizes the reviewed vision-based explanations for AVs.

D. REINFORCEMENT LEARNING AND IMITATION LEARNING-BASED EXPLANATIONS

Explaining how perceived environmental states are mapped to actions has also recently received attention in the autonomous driving community. In this regard, the field of explainable reinforcement learning (XRL) is a relatively new and emerging research avenue on XAI [111], [112], [113]. Like vision-based explanations, XRL techniques also aim to provide some forms of justifications on chosen actions of a vehicle either via intrinsically interpretable design or post-hoc explanations. One of the early works in this context is the Semantic Predictive Control (SPC) framework [114], where the authors propose a data-efficient policy learning approach that predicts future semantic segmentation and provides visual explanations of a learned policy. The framework concatenates multi-scale intermediate features from RGB with tiled actions. The joined modules are then fed into the multi-scale prediction model that predicts future features. Finally, in the last part of the pipeline, the information prediction module inputs the latent feature representation and outputs driving signals alongside the semantic segmentation of the scene.

Chen et al. [115] introduce a sequential *latent environment model* learned with RL and a probabilistic graphical model-based approach interpreting autonomous cars' actions via a bird-eye mask. They use video cameras and LIDAR images as input in the CARLA simulator [116]. For the purpose of interpretability of actions and explainability of a learned policy, they generate a bird-eye mask (i.e., Figure 6). Their model outperforms the used baseline models - DQN, DDPG, TD3, and SAC. Similarly, Wang et al. [120] propose an interpretable end-to-end vision-based motion planning (IVMP) to interpret the underlying actions of a self-driving vehicle. They use semantic maps of a bird-view space in order to plan the motion trajectories of an autonomous car. Moreover, the IVMP approach uses an optical flow distillation network that can improve the real-time performance of the network. The experiments conducted on the nuScenes dataset [134] show the superiority of the proposal over modern approaches in semantic map segmentation and imitation of human drivers. In another probabilistic decision-making model, Wang et al. [121] approach lane merging task as a dynamic process and integrate internal states into joint Hidden Markov Model (HMM) and Gaussian Mixture Regression (GMM). The experiments conducted on the INTERACTION dataset [135] demonstrate the efficiency of the proposed technique and show that merging at highway on-ramps can be delineated by three interpretable internal states in terms of the absolute speed of a vehicle while merging.

Rjoub et al. [122] have shown that federated deep RL combined with XAI can lead to trusted autonomous driving. They use a federated learning approach for decision-making and leverage edge computing that enables different devices to train an ML model in a collaborative manner. The model is first developed on the parameter server and further broadcasted to other devices. Then, global ML methods intake updates from these devices regularly, and the process continues until the model performs well enough on driving tasks. Yang et al. [130] have also shown that an intrinsically interpretable RL agent for autonomous driving can be achieved via reward consistency with the aim of resolving the gradient disconnection in reward-action mapping.

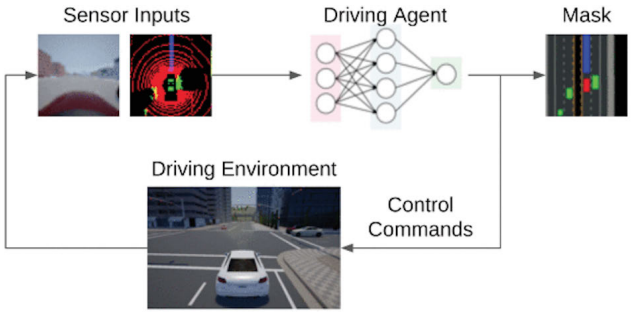

FIGURE 6. RL-based interpretable end-to-end autonomous driving via a bird-eye mask. Credit: [115].

Finally, within the exploration via interaction context, a few studies have employed various forms of imitation learning (IL) techniques for explainable autonomous driving. Cultrera et al. [118] present conditional imitation learning with an end-to-end visual attention model, which identifies those parts of images that have a higher influence on predictions. They test their architecture in the CARLA simulator on four tasks - go straight, turn left, turn right, and follow the lane. Their ablation study focused on box type importance and fixed grid analysis to get an attention map on the images shows that integrated imitation learning and attention model enables a car to drive safely and perform relevant maneuvers in real time.

Leveraging vision for interpretability of an agent's actions, Teng et al. [124] propose a hierarchical interpretable IL (HIIL) technique that unifies bird eye view (BEV) mask with the steering angle to perform actions in complex situations as an end-to-end autonomous driving pipeline. They construct their method as a two-phase task: In the first phase, the pre-trained BEV model is used to interpret the driving environment. Then, IL takes the latent features of the BEV mask from the first phase and combines them with a steering angle acquired through the Pure-Pursuit algorithm. The experiment performed in the CARLA simulator shows that the proposed method enhances the interpretability and robustness of driving in various circumstances.

Moreover, Renz et al. [123] introduce PlanT, a rigorous IL-based approach that uses transformers for planning. PlanT is able to explain its action decision by recognizing the most important object in its driving segment and outperforms state-of-the-art work in CARLA's Longest6 Benchmark by 10 points (See [143] for a visual demonstration). In a more

TABLE 2. Studies on reinforcement learning and imitation learning-based explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Pan et al., [114], 2019	Semantic predictive control for explainable policy learning	LSTM, DDPG-SEG, DLA, model-based RL	Visual	AV developers
Bansal et al., [117], 2019	Leveraging concept bottlenecks as visual features for predicting control command and explanations of vehicle and human behavior	ChauffeurNet, AgentRNN, Perception-RNN, IL	Visual	AV developers
Cultrera et al., [118], 2020	Explaining autonomous driving by learning end-to-end visual attention	CNN, IL	Visual	AV developers
Chen et al., [115], 2021	Interpretable end-to-end autonomous driving with latent deep reinforcement learning	MaxEnt RL, DQN, DDPG, TD3 and SAC	Visual	AV developers
Schmidt et al., [119], 2021	Interpretable and verifiably RL for autonomoous driving learning	SafeVIPER, PPO	Visual	AV developers
Wang et al., [120], 2021	Learning interpretable end-to-end vision-based motion planning with optical flow distillation	IVMP, Optical flow	Visual	AV developers
Wang et al., [121], 2021	Uncovering interpretable internal states of merging tasks at highway on-ramps for autonomous driving decision-making	GMR, HMM	Visual	AV developers
Rjoub et al., [122], 2022	XAI-based federated deep RL for autonomous driving	DQN, DQN-XAI, SHAP	Visual	AV developers
Renz et al., [123], 2022	Explainable planning for autonomous driving	BERT, GRU, IL	Visual	AV developers
Teng et al., [124], 2022	Interpretable imitation learning for end-to-end autonomous driving	Bird's Eye View model, IL	Visual	AV developers
Hejase et al., [125], 2022	Interpretable state representation for deep RL in autonomous driving	DDQN	Visual	AV developers
Cultrera et al., [126], 2023	Visual attention and end-to-end trainable region proposals for explainable autonomous driving	IL, Region Proposal Networks, Spatial Transformers Network	Visual	AV developers
Shao et al., [127], 2023	Interpretable sensor fusion transformer for safe autonomous driving	InterFuser, IL, ResNet, GRU	Visual	AV developers
Paleja et al., [128], 2023	Interpretable continuous control trees for autonomous driving	Differentiable Decision Trees, SAC	Visual	AV developers
Kenny et al., [129], 2023	Interpretable Deep RL with Human-Friendly Prototypes for autonomous driving	PW-Net, PPO, TD3	Visual	AV developers
Yang et al., [130], 2023	Reward consistency for interpretable feature discovery for autonomous driving	PPO	Visual	AV developers
Lu et al., [131], 2024	Human-like cognitive maps for enhancing interpretability of autonomous driving	Successor Representations, Cognitive Potential Field	Visual	AV developers
Liu et al., [132], 2024	Interpretable generative adversarial IL for autonomous driving	IL, Signal Temporal Logic	Visual	AV developers
Wang and Aouf, [133], 2024	Explainable deep adversarial RL for robust autonomous driving	PPO	Visual	AV developers

recent work in this context, Liu et al. [132] show that combining Signal Temporal Logic with generative IL is also an effective approach for interpretable policy for autonomous driving. Overall, while building interpretable by-design RL or IL agents for autonomous driving is a challenging task, employing external methods, such as logic and vision, can help achieve explainability for the agent's actions. Table 2 summarizes RL and IL-based explanations for AVs.

E. FEATURE IMPORTANCE-BASED EXPLANATIONS

Being inherently interpretable and easier to understand a prediction of a model, feature importance scores, particularly decision-tree-based explanations and SHAP values [144], have also been investigated in autonomous driving (Table 3). Decision trees have been proven to describe the rationale semantically for each prediction made by a CNN architecture [145]. Omeiza et al. [136] use decision trees as a *tree-based* representation that generates scenario-based explanations of different types by mapping observations to actions in accordance with traffic rules. They use human evaluation in a variety of driving scenarios and generate Why, Why Not, What If, and What explanations for driving situations and empirically prove that the approach is effective for the intelligibility and accountability goals of automated vehicles.

Brewitt et al. [137] introduce Goal Recognition with Interpretable Trees (GRIT), a framework that uses decision trees trained from the trajectory data of a self-driving car. The framework, tested on fixed-frame scenarios, is proven empirically verifiable for goal recognition using a satisfiability modulo theories (SMT) solver [146].

Cui et al. [138] use Random Forest for the interpretability purpose on the autonomous car-following task. They employ deep RL for the decision-making of an autonomous car and employ SHAP values to simplify the feature space. Once the agent generates state-action pairs, Random Forest is applied to these pairs and experimental results show the approach works effectively to explain behavior for the designated car-following task. In a recent study, Random Forest has also been proven to detect misbehaving vehicles in Vehicular Adhoc Networks (VANET) in Mankodiya et al.'s work [70]. Finally, feature importance scores have been

TABLE 3. Studies on feature importance-based explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Omeiza et al., [136], 2021	Generating tree-based explanations with and without causal attributions	Tree-based representation / User study	Textual	All groups
Brewitt et al., [137], 2021	Interpretable and verifiable goal recognition with learned decision trees for autonomous driving	Decision Tree	Visual and Textual	AV developers
Mankondiya et al., [70], 2021	XAI for trust management in autonomous vehicles	Random Forest, Decision Tree, AdaBoost	Visual	AV developers
Cui et al., [138], 2022	Interpretation framework for autonomous driving	Random Forest, SHAP	Visual	AV developers
Onyekpe et al., [139], 2022	AV positioning using SHAP	SHAP, WhONet	Visual and Textual	AV developers
Almalioglu et al., [140], 2022	Vehicle position with deep learning	GRAMME, SHAP	Visual	AV developers
Ayoub et al., [141], 2022	Predicting driver takeover time in conditional automated driving	XGBoost, SHAP	Visual	AV developers
Brewitt et al., [142], 2023	Interpretatable trees for goal recognition in autonomous driving under occlusion	OGRIT, Decision Tree	Visual	AV developers

used for vehicle positioning [139], [140] and predicting driver takeover time [141] in various levels of autonomous driving Thus, being computationally more transparent than traditional deep neural network architectures, decision trees can explain behaviors of a variety of autonomous driving tasks with less computation.

F. LOGIC-BASED EXPLANATIONS

While the interpretability of a deployed autonomous driving control model has been the dominant direction for research, there have also been attempts to verify the safety of self-driving vehicles with logical reasoning. In this regard, Corso and Kochenderfer [148] present a technique to identify interpretable failures of autonomous cars. They use *signal temporal logic* expressions to describe failure cases of an autonomous car in an unprotected left turn and pedestrian crossing scenarios. For this purpose, the authors use genetic programming to optimize signal temporal logic expressions that acquire disturbances trajectories, causing a vehicle to fail in its decisive action. The experimental results show that the proposed approach is effective in interpreting the safety validation of a car.

Suchan et al. [147] have developed an *answer set programming*-based abductive reasoning framework for online sensemaking for perception and control tasks. In its essence, the framework integrates knowledge representation and computer vision in an online manner to explain the dynamics of traffic scenes, particularly occlusion scenarios. The authors demonstrate their method's explainability and commonsensical value with empirical study collected on the KITTI MOD [151] dataset and the MOT benchmark [152]. Another experimental study leveraging the concept of answer set programming has been carried out by Kothawade et al. [150]: they introduce AUTO-DISCERN, a system that incorporates common sense reasoning with answer set programming to automate explainable decision-making for self-driving vehicles. They test their rules and show AUTO-DISCERN's credibility in real-world scenarios, such as lane changing and right turn operations, from the KITTI dataset. Table 4 summarizes logic-based explanations for AVs.

G. USER STUDY-BASED EXPLANATIONS

Some investigations involve users in case studies to understand the effective strategies for explanation generation in autonomous driving tasks. The key idea of a user study is that getting people's input in designated driving tasks can help improve the adequacy and quality of explanations in autonomous driving. Wiegand et al. [153] perform a user study that identifies a mental model of users for determining an effective practical implementation of an explanation interface. The main research question here is to understand what components need to be visualized in a vehicle so the user can comprehend the decisions of self-driving vehicles. The study discloses that combining an expert mental model with a user mental model as a target mental model enhances the situation awareness of the drivers. Furthermore, Wiegand et al. [154] investigate situations in which explanations are needed and methods pertinent to these situations. They spot seventeen scenarios where a self-driving vehicle behaves unexpectedly. Twenty-six participants are selected to validate these situations in the CarMaker driving simulator to provide insights into drivers' need for explanations. As a result of the user study, the authors identify six groups to highlight the primary concerns of drivers with these unexpected behaviors, namely emotion and evaluation, interpretation and reason, the capability of a self-driving car, interaction, driving forecasting, and request times for explanations.

Wang et al. [66] propose an approach that enables a human driver to provide *scene forecasting* to an intelligent driving system using a purposeful gaze. They develop a graphical user interface to understand the effect of human drivers on the prediction and control of an intelligent vehicle. A simulator is used to test and verify three driving situations where a human driver's input can improve safety of self-driving.

Apart from these works, Schneider et al. involve human participants in their empirical studies to understand the role of explanations for the public acceptance of AVs [93], [155]. They explore the role of explainability-supplied UX in AVs, provide driving-related explanations to end users with different methods, such as textual, visual, and lighting techniques, and conclude that providing context-aware explanations on

TABLE 4. Studies on logic-based explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Suchan et al., [147], 2019	An answer set programming-based abductive reasoning for visual sensemaking	Answer set programming, YOLOv3, SSD, Faster R-CNN	Visual	AV developers
Corso and Kochenderfer [148], 2020	Interpretable safety validation for autonomous driving	Signal temporal logic	Textual	AV developers
DeCastro et al., [149], 2020	Interpreting policies via signal temporal logic for autonomous driving	Signal temporal logic, LSTM, CVAE	Visual	AV developers
Kothawade et al, [150], 2021	Explainable autonomous driving using commonsense reasoning	ASP, s(CASP)	Textual	Road users

autonomous driving actions increases users' trust in this technology. Their subsequent study also confirms that driving explanations can help mitigate the negative impact of AVs failures on users [156]. Finally, Kim et al.'s user study [157] confirms that humans do not need explanations seamlessly, and presenting explanations only in critical driving conditions is preferred to enjoy the trip with an autonomous car and prevent information overload. Table 5 summarizes user studies on XAI for AVs.

H. LARGE LANGUAGE MODELS AND VISION-LANGUAGE MODELS-BASED EXPLANATIONS

Finally, while the preliminary studies and further works on explainable autonomous driving primarily focused on a combination of various AI techniques revisited above, large language models (LLMs) and vision-language models (VLMs) have recently emerged as a novel paradigm for interpreting AV decisions and describing traffic scenes. Built on top of Foundation Models, such as GPT [159] and BERT [160], there has been significant progress on building both domain-agnostic and domain-specific LLMs (e.g., GPT-3 [161], GPT-4 [162], LLAMA [163], LLAMA-2 [164], Vicuna [165], Alpaca [166], Claude [167]) and VLMs (e.g., Flamingo [168], LLaVA [169], PaLM-E [170], Video-LLaVA [171], Video-LLaMA [172], Gemini [173], Claude 3 [174]). In this sense, there have been tremendous efforts to build language and vision-language models on top of these base models for interpretable autonomous driving. Overall, based on the recent trend in leveraging these large models for the interpretability purpose in autonomous driving, we observe the following directions:

1) PRESENTING LIVE NATURAL LANGUAGE EXPLANATIONS DURING THE TRIP

The promising work in this context is Wayve's LINGO-1 [175] and LINGO-2 [176] architectures. LINGO is a vision-language-action (VLAM) model that provides live natural language explanations for describing a vehicle's chosen actions in end-to-end autonomous driving. Trained on diverse multimodal (vision and language) datasets, LINGO can describe action decisions and causal factors inducing these actions. The advantage of the LINGO architecture is that its explanations are concise, informative, and reflect temporal changes in the driving environment. The Wayve team has also achieved live linguistic explanations for autonomous driving in a simulation environment [177].

2) VIDEO QUESTION ANSWERING AS A REASONING TECHNIQUE

An essential characteristic of modern AVs is the consideration of human factors in the design and development of this technology and having effective human-machine alignment for trustworthy autonomous driving. In this sense, it is crucial that users-in-the-loop have some form of interaction with AVs during a journey. Motivated by this concept, some recent works have approached conversational user interface between people on board and AVs as a Video Question Answering (VideoQA) task [178], [179], [180]. Asking questions about the behavior of an autonomous system is a part of our intuition, and in the context of AVs, getting answers to traffic-related situations and autonomous car action-related questions can help users have a comfortable and reliable journey. Other practical applications of LLMs and VLMs are interpretable motion planning [181], chain-of-thought-based reasoning for control and decision-making [182], [183], action justification with a control signal [184], predicting the intent of other traffic actors via visual reasoning and cues [185], and understanding the role of video transformer based-explanations on safety of autonomous driving [186] (see Table 6 for more details on LLM and VLM-based explanations).

Overall, as an emerging AI technology, LLMs and VLMs have tremendously benefitted AVs from the interpretability aspect, as described in the above studies. However, it is also worthwhile to mention that there are still spaces for improvement of these models as fictitiously generated explanations (e.g., *hallucinations*) may have serious safety implications and high-stakes consequences for self-driving actions and human life. We describe such caveats and potential solutions in Section VI as future work.

A high-level overview of all these studies indicates driving explanations are generally multi-modal, context-dependent, and task-specific, justifying action decisions of AVs. Moreover, end-to-end learning has become even more popular for highly autonomous decision-making owing to the combination of powerful deep-learning approaches and overall safety benefits. Based on the insights from the state of the art, we can define explainable autonomous driving as follows:

> *Explainable autonomous driving is a self-driving approach powered by a compendium of AI techniques 1) ensuring an acceptable level of safety*

TABLE 5. User study-based explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Wiegand et al., [153], 2019	Explaining driving behavior of autonomous cars	User study	Textual	Backup drivers
Wiegand et al., [154], 2020	Understanding situations that a driver needs explanations	User study	Visual	All groups
Wang et al., [66], 2021	Enhancing automated driving with human foresight	User study	Visual	Backup drivers
Schneider et al., [155], 2021	UX for transparency in autonomous driving	UEQ-S, AVAM (User study)	Visual, Textual, Light	All groups
Schneider et al., [93], 2021	Increasing UX through different feedback modalities	UEQ-S (User study)	Visual, Textual, Audio, Light, Vibration	All groups
Shen et al., [158], 2022	Identifying which scenarios need explanations in autonomous driving	Friedman test, Pearson correlation, Point-Biserial Correlation	Visual	Road users
Schneider et al., [156], 2023	The role explanatory information in failure situations in highly autonomous driving	UEQ-S, AVAM	Visual, Textual	All groups
Kim et al., [157], 2023	Timing perspective and mode of explanations for road users in autonomous driving	GradCam, Head-mounted display, Windshield display	Visual	Road users

for a vehicle's real-time decisions, 2) providing explanatory information on the action decisions in critical traffic scenarios in a timely manner for transparency, and 3) obeying all traffic rules established by the legal entities and regulators.

Driven by this definition and the state-of-the-art works in the above sections, we present a conceptual XAI framework for end-to-end autonomous driving aligned with industrial trends and show the necessary components, process steps, and critical challenges toward achieving regulatory-compliant AVs in the next generation.

V. AN XAI FRAMEWORK: INTEGRATING END-TO-END CONTROL, SAFETY, AND EXPLANATIONS

We present a general framework in which methods for developing XAI, end-to-end learning, and safety components are combined to inform processes of regulatory principles. Each of these components has a concrete role in our framework. In our recent study [191], we have covered a brief description of end-to-end learning for AVs. We extend the scope of that work and describe the essential elements of end-to-end autonomous driving, and the role of and potential challenges with explanations in such a setting. We describe these individual components as follows:

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

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

and a compendium of actions

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

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

$$C : E \mapsto A.$$

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

$$e \in E$$

to a relevant action

$$a \in A.$$

The most prevalent learning paradigms for end-to-end autonomous driving are RL and IL [17]. Furthermore, differentiable learning has also recently emerged as an end-to-end driving architecture: While the planning component is prioritized, this learning pipeline optimizes several modules of the entire driving architecture (e.g., [18]). Overall, as described in Section II, the end-to-end learning pipeline uses a single deep neural network as a unified task to map the sensor model of the world to real-time control commands of AVs.

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

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

This requirement could be as pragmatic as some inspection of individual vehicle safety (for example, verifying basic safety functions of an individual vehicle for re-licensing). That said, this concept should be considered as a thorough compliance testing of *eeC* components from vehicle manufacturers to certify their public safety under international and/or national transportation guidelines such as [33] and [38]. The general principles for acceptable functional safety of road vehicles are defined by the ISO 26262 standard [44]. According to this standard, there should be a safety certification development with evidence-based rationales: the vehicle should be able to meet the established functional safety requirement in its operational context. Part 6 of the ISO 26262 standard [192] is dedicated to end-product development for automotive applications within the software level. This guideline includes the design, development, testing, and verification of software systems in automotive applications.

TABLE 6. Studies on large language models and vision-language models-based explanations for AVs.

Study	Task	Algorithms/Methods	Delivery format	Target audience
Dewangan et al., [185], 2023	Language-augmented Bird's-eye View Maps for Autonomous Driving	GPT-4, GRIT	Visual, Textual	All groups
Xu et al., [178], 2023	VQA and natural language-based explanations for autonomous driving	LLAMA 2, CLIP	Visual, Textual	All groups
Marcu et al., [179], 2023	Video question answering for autonomous driving	Vicuna-1.5-7B, GPT-4	Textual	All groups
Chen et al., [177], 2023	Improving context understanding in autonomous driving with object-level vector modalities and LLM	GPT 3.5, PPO	Visual, Textual	All groups
Fu et al., [187], 2023	Understanding traffic situations in a closed loop	GPT 3.5	Textual	All groups
Mao et al., [181], 2023	Interpretable motion planning as language modeling	GPT 3.5	Textual	Road users
Sha et al., [182], 2023	LLM as a decision-maker in complex driving scenarios	ChatGPT, MPC	Visual, Textual	AV developers
Wayve Team [175], 2023	Providing live explanations in natural language	Integrated vision, language and action architecture	Textual	All groups
Nie et al., [188], 2023	Interpretable reasoning in complex driving situations in autonomous driving	GPT-4, MLP, ViT-G/14	Textual	All groups
Park et al., [180], 2024	Video question answering for traffic scene understanding	Video-LLAMA, GPT-4	Textual	All groups
Wen et al., [183], 2024	LLM-based knowledge-driven approach for interpretable autonomous driving	Out-of-box LLM	Visual, Textual	AV developers
Yuan et al., [184], 2024	Retrieval-augmented VLM for explainable autonomous driving	LanguageBind, MLP, Vicuna-1.5-7B	Textual	All groups
Chi et al., [189], 2024	GPT-aided explainable decisions for autonomous vehicles	GPT, Graph of Thoughts	Textual	All groups
Atakishiyev et al., [186], 2024	Robustness of a transformer-based VideoQA model against human-adversarial questions and its safety implications for self-driving	Video-LLaVA	Textual	All groups
Duan et al., [190], 2024	Unifying imitation learning with LLMs to enhance safety of end-to-end driving	Vicuna LLM, Swin transformer	Textual	All groups

Based on these standards, there seem to be two fundamental approaches to confirming regulatory compliance, which we label confirmation of compliance by ''simulation,'' and confirmation of compliance by ''verification,''. These steps are aligned with 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 satisfactory. This approach is perhaps the most familiar, as it arises naturally from an engineering development trajectory, where the accuracy of simulators determines the quality of compliance (e.g., [193]). The confidence of the established compliance is a function of the accuracy and coverage of the simulation. However, this compliance process can be very expensive and prone to safety gaps, especially when consensus on the properties and scope of a simulation is difficult to achieve. Thus, in general, the simulation part can be considered a ''driving school'' for AVs: The designed and developed learning software system should be tested rigorously in this phase before such an autonomous system, as a holistic architecture, is deployed to a vehicle in the physical environment and on real roads.

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

In addition to safety assurance, another critical requirement of AVs is their ability to defend against adversarial attacks. The ISO/SAE 21434 standard has defined guidelines for cybersecurity risk management for road vehicles, and AVs must also comply with these requirements [195]. As AVs increasingly rely on their automation ability, it is vital that ML software of an intelligent driving system and built-in interfaces can detect and defend against potential cyber-attacks of the broad spectrum, such as electronic control units (ECU) attacks, in-vehicle network attacks, and automotive key-related attacks [196], [197], [198].

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

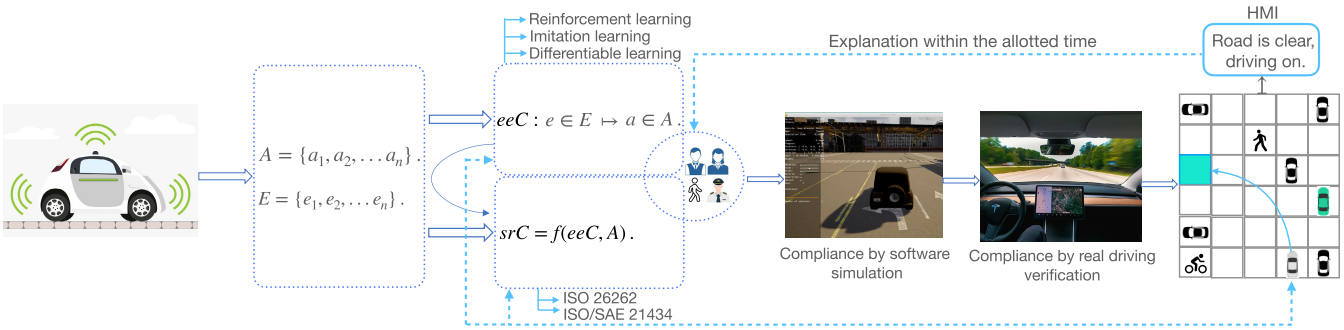

FIGURE 7. A diagram of the proposed explainable end-to-end autonomous driving framework.

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

A. TEMPORAL GRANULARITY AND CONVEYANCE OF EXPLANATIONS

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

1) TIMING MECHANISM OF EXPLANATIONS

Delivering timely explanations can help human drivers/passengers react to emergent situations, such as takeover requests, appropriately and prevent a potential danger in the vicinity. According to Koo et al.'s study [199], it is favorable to convey explanations before a driving event is about to happen. This concept has further been validated by Haspiel et al.'s user study, and human judgment shows that explanations should be delivered *before* action is *decided* rather than *after* it is *performed* [200]. This judgment makes sense as on-time communication of explanations can bring situation awareness for people on board and enable them to monitor an autonomous car's subsequent action. If the action to be performed soon is hazardous, a human driver or passenger can manually intervene in the situation with such explanations and prevent a potential danger ahead.

2) THE IMPACT OF LEAD TIME ON THE SAFE TRANSITION FROM AN AUTOMATED MODE TO A HUMAN TAKEOVER

Another important criterion is determining the amount of time needed to alert human actors for a takeover request. In the user study measuring the impact of 4 s vs. 7 s as the lead time on takeover alert, Huang and Pitts [201] show that a shorter lead time leads to a faster transition to human-controlled mode but also lacks the quality of takeover as lack of time may be stressful for a human actor in such situations. A similar conclusion has been acquired by Mok et al. [202] in the case of 2, 5, and 8 s transition times. Wan and Xu [203] have further verified that an insufficient amount of lead time, such as 3 s, results in an impaired takeover performance, and drivers perform better when enough time, such as more than 10 s, is allotted for takeover requests. In general, it can be concluded that lead time for explaining emergent situations to a human and transitioning control should happen within a few seconds, while for non-critical situations, such as post-trip analysis, the amount of time may be as long as needed.

3) ALL-TIME OR ONLY NECESSARY EXPLANATIONS?

It is also important to consider that humans need to enjoy their trips with AVs and get information from a vehicle only when it is necessary. This aspect also applies to the delivery of explanatory information to end users. When the passengers/human drivers are provided with tons of information during the trip, it can lead to *mental overload* for them [154]. Consequently, it is generally favored that driving decisions and traffic scenes may be described to humans on board when traffic conditions are critical, and people need to be alerted.

It is also noteworthy to specify that AVs must be equipped with need-based HMIs to deliver explanations. There are some challenges with effective automotive HMI design. First, people may have different choices or preferences for HMI (i.e., display monitor, alert interface, etc.). Furthermore, users' various cognitive and functional abilities must be a crucial factor in the design of user interfaces [82]. For instance, people with visual or hearing impairments may need a customized HMI. Hence, automotive manufacturers must consider the diversity of users, contemplate the timing

perspective of HMI explanations in line with relevant actions, and reach a consensus on the best practices with effective HMI design for AVs [204].

Based on the mentioned process steps and crucial elements, we see that achieving the interpretability of self-driving models is challenging, necessitating integration of those steps and cooperation between users and AVs. Consequently, while we argue that transparent and highly autonomous driving is feasible, human factors must be a vital consideration in the design and development of such systems. A simple graphical illustration of our proposed framework with its elements can be seen in Figure 7. It is also noteworthy to specify that previous XAI frameworks for autonomous driving, such as in [46] and [94], focus on the *modular* pipeline, while our framework, to the best of our knowledge, is the first one proposed for *end-to-end* autonomous driving, aligning with current trends in the automotive industry.

In the next section, we envision the future of AV research in the realm of modern AI technologies and safety and explainability based on current industrial trends and list some potential challenges toward this goal.

VI. TOWARD AV 2.0: UNIFYING VISION, LANGUAGE, AND ACTION WITHIN EMBODIED AI FOR SAFE AND EXPLAINABLE END-TO-END AUTONOMOUS DRIVING

While the above subsections primarily describe the potential XAI approaches from the perspective of specific components, we also need to consider AVs' learning software as a holistic driving system. Three decades of research, starting with ALVINN in 1988 [14] and further succeeding with the DARPA Grand Challenge [205], have achieved significant milestones with traditional AI software. However, recent breakthroughs in Foundation Models in terms of LLMs and VLMs motivate a transition to next-generation AVs. This generation of AVs has been referred to as *AV2.0* by industry professionals [206], [207], [208]. The proposal is that the availability of integrated sensor suites, computational resources (i.e., GPU, TPU), and deep learning approaches can help AVs navigate via an end-to-end approach through adaptive learning, scaling, and generalization in complex driving environments. The ability to learn continually through interaction with the environment rather than relying on static datasets has resulted in the emergence of a new direction, labeled as "Embodied AI" [209], [210], and AV2.0 research can move forward with such a learning approach. Effectively unifying vision, language, and action within Embodied AI can enable an autonomous car to navigate, interpret, and describe its high-level decisions in real-time. However, safety and explainability components of an end-end self-driving architecture must overcome fundamental challenges in AI described below:

Safety: The established guideline on core problems with AI safety [211] underscores five crucial considerations: avoiding negative side effects, avoiding reward hacking, scalable oversight, safe exploration, and robustness to distributional shift. We analyze the implications of these problems for end-to-end autonomous driving as follows:

- **Avoiding negative side effects**: Autonomous driving is primarily associated with the ability of a self-driving car to avoid accidents and maintain a safe distance from stationary and dynamic objects along the planned motion trajectory. However, the scope of the problem is not limited to this feature. Consider a scenario where an autonomous car interacts with two other vehicles, V1 and V2, at a specific moment. While aiming to make safe temporal decisions by itself, the autonomous car must also ensure that it does not implicitly enable V1 and V2 to cause an accident at that road segment as a part of vehicle-to-vehicle (V2V) communication. According to [211], a potential solution to this problem could be to leverage cooperative Inverse RL [212], where an autonomous system can cooperate with humans, and a human actor can always shut down the autonomous system in case such a system exhibits undesirable behavior. In the context of autonomous driving, this nuance can be related to an AV's communication with a human-operated vehicle or other remote operator monitoring an AV's overall driving safety. One of the prominent methods in this context is Sympathetic Cooperative Driving or SymCoDrive paradigm [213], which trains agents not only to achieve safe driving for themselves but also for human-controlled vehicles by promoting altruistic driving behavior in cooperative autonomous driving. As the deployment of AVs on roads is a gradual process, synergy with human-operated vehicles is a viable approach for socially aware and safe navigation.
- **Avoiding reward hacking**: Can we ensure that the end-to-end driving system does not shape its dynamic reward function according to what it sees in less dynamic environments and still apply that reward shaping while transitioning to highly dynamic environments? Particularly, as an embodied AI agent with adaptive learning and generalization ability in unseen environments, reward formulation must account for long horizons ahead and should not adjust its goals for short-term safe driving behavior. This topic has recently been well-investigated by Knox et al. [214]. They propose that flaws in reward shaping for RL-controlled autonomous driving can be identified by *eight sanity checks*: unsafe reward shaping, potential mismatch between people and reward function's preferences, undesired risk tolerance via indifference points, learnable loopholes, missing attributes, redundant attributes, and trial-and-error reward design. The study discloses that such sanity checks can capture flaws in reward shaping for autonomous driving that can also exist in reward shaping for other tasks.
- **Scalable oversight**: Can humans measure whether AVs perform at a human level or better in general in all driving situations, where in specific moments, evaluating the driving behavior of end-to-end driving may be difficult for humans due to some reasons? While being outside of human override, temporarily (i.e., refer to the Molly

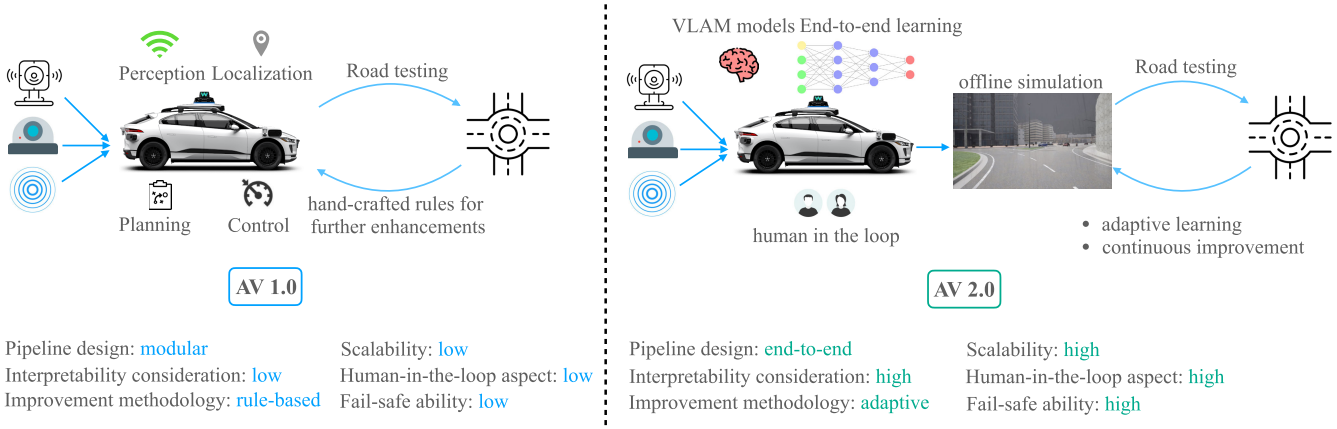

FIGURE 8. Our approach to AV2.0 vs AV1.0, and potential advantages of AV2.0 over AV1.0 in terms of its AI software stack, safety and explainability. The image of the autonomous vehicle has been taken from Waymo's media resources.

problem [215]), for various reasons, can we trust that AVs will behave safely at that moment? Amodei et al. [211] report that a potential solution to this problem may be semi-supervised RL: an agent can see its reward on a small subset of episodes or times steps. While rewards from all episodes are used to evaluate the agent's performance, the agent can only use that subset of rewards to optimize its performance under this setting.

- **Safe exploration**: Can an AV always make safe decisions when it has a binary choice of actions in a specific time interval? For example, an autonomous car may change its predefined route due to traffic congestion; however, the alternative route may have dangerous potholes or other damaged infrastructure that may lead to risky driving while attempting to save time on the trip.
- **Robustness to distributional shift**: A well-known problem with AVs is making the transition from a simulation environment to real roads. For instance, RL-based end-to-end autonomous driving with an impressive performance in a simulation environment may not have the same performance when deployed to a physical autonomous car. Filos et al. [216] have investigated this topic and proposed *robust imitative planning*, a technique for epistemic uncertainty-aware planning. The key idea is that in case the model has great uncertainty in suggesting a safe course of action, the model can achieve sample-efficient online adaptation by querying the expert driver for feedback. Through several experiments and state-of-the-art results, the authors also release CARNOVEL, a benchmark for evaluating the robustness of driving agents with distribution shifts. Such a benchmark may be a significant part of a robust solution for addressing out-of-distribution scenarios.

These problems reflect a broad spectrum of potential safety issues with end-to-end AVs. However, we argue that the proposal misses yet another essential concept, namely *fail-safe* ability. This concept has been investigated in some recent work [217], [218], [219]; however, the recent propositions of the next-generation AVs [206], [207], [208] do not explicitly consider this functionality as an integral component of this technology. Human drivers often have a rest once they feel tired on long trips, and a short rest may help them feel mentally/physically better in the next phase of their driving. The same example can applied to AVs as well. Due to internal reasons (e.g., temporary system malfunction) or external factors (e.g., extremely adverse weather conditions), AVs may need to pause their trip temporarily and prevent further high-stakes consequences ahead. Such capability should not be considered a limitation of AVs; on the contrary, it is an optimal design strategy that foresees potential issues due to *any* factors and makes AVs behave safely by directing them to "have a short rest."

Explainability: The reviewed studies in Section IV show a significant milestone in the explainability of self-driving systems. However, there are still significant gaps and challenges to achieving accurate and timely explanations in all phases of trips. For instance, as of September 2023, it is reported that LINGO-1 exhibits roughly 60% performance in its linguistic and VQA-based explanations compared to human-level performance [175].

Apart from informational content, another critical aspect deserving attention is the timing perspective of such explanations: the lead time for emergent scenarios, perhaps using extensive scenario-based evaluations or case-based reasoning, must be engineered appropriately. Furthermore, a well-known problem with large pre-trained models, hallucinations, is another challenge in explanation delivery. Particularly, in QA models, the model must generate a response based on the joint question and scene-based semantics rather than being influenced by the question itself, such as in the case of adversarial queries. We have recently performed an empirical study [186] on the latter and shown that even advanced VLMs can fail to detect the language bias in QA models and present incorrect explanations in case of human adversarial questions. This issue, in turn, may damage user trust and can also have negative safety implications for self-driving. So, we argue that large pre-trained models' construction mechanisms can be adjusted and regulated with common sense and human-defined concepts, as also posited

by Kenny and Shah [220]. Hence, designing robust QA models deserves more attention to enable meaningful and trustworthy dialogues between users and AVs. These features are key for achieving effective human-AI alignment [221], [222], [223], trust [7], [224], and public acceptance [46], [223], [225] with AV2.0 within the principles of regulated AI [10]. Figure 8 describes our perspective on AV2.0 and its difference from AV1.0, complementing goals of the XAI framework for end-to-end autonomous driving in Figure 7 with a modern approach.

VII. CONCLUSION

In this paper, we have presented a systematic overview of state-of-the-art investigations, emerging paradigms, and a future perspective of XAI approaches for autonomous driving. Insights from these studies reveal the existing gaps, and we have proposed a conceptual framework for explainable autonomous driving by incorporating missing pieces. The key idea is that AVs need to achieve regulatory-compliant operational safety and explainability in real-time decisions with their increasing automation ability. Together with a detailed overview of the state of the art in XAI-based autonomous driving, our work contributes as a *cause-effect-solution* perspective. We elaborate on the notion of *cause* by identifying current gaps, concerns, and a variety of issues specified while denoting *effect* through the public reluctance on the use of autonomous driving at a broader level. We provide a *solution* through the proposed framework and a set of promising XAI approaches for a future direction. This paper can benefit automotive researchers and practitioners in understanding the emerging paradigms and industrial trends in XAI approaches for autonomous driving and help achieve responsible, trustworthy and publicly acceptable next-generation AVs.

REFERENCES

[1] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Tech. Rep., 2015.
[2] R. Lanctot, "Accelerating the future: The economic impact of the emerging passenger economy," *Strategy Analytics*, vol. 5, pp. 1–20, Oct. 2017.
[3] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[4] 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," *Saf. Sci.*, vol. 120, pp. 117–128, Dec. 2019.
[5] 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.
[6] (2024). *Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View*. Accessed: Mar. 10, 2024. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf
[7] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.
[8] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, "Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems," *Transp. Res. Part C, Emerg. Technol.*, vol. 156, Nov. 2023, Art. no. 104358.
[9] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110273.
[10] Council EU. (2024). *Artificial Intelligence (AI) Act: Council Gives Final Green Light to the First Worldwide Rules on AI*. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/
[11] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Syst. Appl.*, vol. 165, Aug. 2021, Art. no. 113816.
[12] S. Campbell, N. O'Mahony, L. Krpalcova, D. Riordan, J. Walsh, A. Murphy, and C. Ryan, "Sensor technology in autonomous vehicles: A review," in *Proc. 29th Irish Signals Syst. Conf. (ISSC)*, Jun. 2018, pp. 1–4.
[13] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, Mar. 2021.
[14] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 1988, pp. 1–28.
[15] Daimler media. *Autonomous Concept Car Smart Vision EQ Fortwo: Welcome to the Future of Car Sharing*. Accessed: Mar. 10, 2024. [Online]. Available: https://media.mbusa.com/releases/release-80848dccd3f3680a764667ad530987e9-autonomous-concept-car-smart-vision-eq-fortwo
[16] J. Shuttleworth. (2019). *SAE Standard News: J3016 Automated-Driving Graphic Update*. [Online]. Available: https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic
[17] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," 2023, *arXiv:2306.16927*.
[18] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17853–17862.
[19] J. Araluce, L. M. Bergasa, M. Ocaña, Á. Llamazares, and E. López-Guillén, "Leveraging driver attention for an end-to-end explainable decision-making from frontal images," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 1–12, Jul. 2024.
[20] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A survey of end-to-end driving: Architectures and training methods," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1364–1384, Apr. 2022.
[21] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial–temporal feature learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 533–549.
[22] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, and Z. Ding, "Inaction: Interpretable action decision making for autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 370–387.
[23] 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, Aug. 2019.
[24] J. Fleetwood, "Public health, ethics, and autonomous vehicles," *Amer. J. Public Health*, vol. 107, no. 4, pp. 532–537, Apr. 2017.
[25] P. Foot, "The problem of abortion and the doctrine of the double effect," *Oxford review*, vol. 5, 1967.
[26] J. J. Thomson, "The trolley problem," *Yale Law J.*, vol. 94, no. 6, p. 5, 1985.
[27] 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, Nov. 2018.
[28] B. Lundgren, "Safety requirements vs. Crashing ethically: What matters most for policies on autonomous vehicles," *AI Soc.*, vol. 36, no. 2, pp. 405–415, Jun. 2021.
[29] J. Harris, "The immoral machine," *Cambridge Quart. Healthcare Ethics*, vol. 29, no. 1, pp. 71–79, Jan. 2020.
[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," *Artif. Intell.*, vol. 279, Feb. 2020, Art. no. 103201.
[31] S. Sohrabi, H. Khreis, and D. Lord, "Impacts of autonomous vehicles on public health: A conceptual model and policy recommendations," *Sustain. Cities Soc.*, vol. 63, Dec. 2020, Art. no. 102457.

[32] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, "Ethical issues in focus by the autonomous vehicles industry," *Transp. Rev.*, vol. 41, no. 5, pp. 556–577, Sep. 2021.
[33] *Regulation EU 2016/679 of the European Parliament and of the Council of 27 Apr. 2016*, GDPR, 2016.
[34] The High-Level Expert Group on AI at the European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. Accessed: Mar. 11, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[35] (2016). *NACTO Policy Statement on Automated Vehicles*. Accessed: Mar. 10, 2024. [Online]. Available: https://nacto.org/wp-content/uploads/2016/06/NACTO-Policy-Automated-Vehicles-201606.pdf
[36] *Federal Automated Vehicles Policy: Accelerating Next Revolution Roadway Safety*, Nat. Highway Traffic Saf. Admin., 2016.
[37] Dept. Transp. (2022). *Occupant Protection for Vehicles With Automated Driving Systems*. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-03/Final-Rule-Occupant-Protection-Amendment-Automated-Vehicles.pdf
[38] *Guidelines for Testing Automated Driving Systems in Canada*, Transport Canada, 2021.
[39] *Act Amending the Road Traffic Act and the Compulsory Insurance Act (Autonomous Driving Act)*, Bundesanzeiger Verlag Board, Westfalen, Germany, 2021.
[40] *Safe Use of Automated Lane Keeping System (ALKS) Summary of Responses and Next Steps*, Dept. for Transp. team, 2021.
[41] *Guidelines for Trials Automated Vehicles Aust*, Nat. Transp. Commission, 2017.
[42] *Outline Systematic Preparations Rel. To Auto. Driving*, Adv. Inf. Telecommun. Netw. Soc., 2017.
[43] *Road Vehicles—Safety of the Intended Functionality*. Accessed: Mar. 10, 2024. [Online]. Available: https://www.iso.org/standard/70939.html
[44] Road vehicles—Functional safety. Accessed: Mar. 10, 2024. [Online]. Available: https://www.iso.org/standard/68383.html
[45] An overview of taxonomy, legislation, regulations, and standards for automated mobility. Accessed: Apr. 8, 2024. [Online]. Available: https://www.apex.ai/post/legislation-standards-taxonomy-overview
[46] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.
[47] U. Ehsan and M. O. Riedl, "Human-centered explainable AI: Towards a reflective sociotechnical approach," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2020, pp. 449–466.
[48] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle," in *Proc. Designing Interact. Syst. Conf.*, Jun. 2021, pp. 1591–1602.
[49] D. Lewis, *Causal Explanation*, 1986.
[50] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.
[51] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[52] *A Vision for Prioritizing Human Well-Being With Artificial Intelligence and Autonomous Systems*, IEEE Global Initiative, 2016.
[53] B. W. Israelsen and N. R. Ahmed, "'Dave. I can assure you. that it's going to be all right' a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," *ACM Comput. Surveys*, vol. 51, no. 6, pp. 1–37, Nov. 2019.
[54] D. Mullen, *Mercedes to Accept Legal Responsibility for Accidents Involving Self-Driving Cars*, 2022. [Online]. Available: https://www.driving.co.uk/news/technology/mercedes-to-accept-legal-responsibility-for-accidents-involving-self-driving-cars/
[55] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, and K. Zieba, "VisualBackProp: Efficient visualization of CNNs," 2016, *arXiv:1611.05418*.
[56] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[57] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 563–578.
[58] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual scene understanding for autonomous driving using semantic segmentation," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019, pp. 285–296.
[59] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8652–8661.
[60] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, "Multi-modal probabilistic prediction of interactive behavior via an interpretable model," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 557–563.
[61] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9520–9529.
[62] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9658–9667.
[63] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? Towards driver-centric risk assessment: Risk object identification via causal inference," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10711–10718.
[64] S. Casas, A. Sadat, and R. Urtasun, "MP3: A unified model to map, perceive, predict and plan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14398–14407.
[65] J. Kim, A. Rohrbach, Z. Akata, S. Moon, T. Misu, Y. Chen, T. Darrell, and J. Canny, "Toward explainable and advisable model for self-driving cars," *Appl. AI Lett.*, vol. 2, no. 4, p. e56, Dec. 2021.
[66] C. Wang, T. H. Weisswange, M. Krüger, and C. B. Wiebel-Herboth, "Human-vehicle cooperation on prediction-level: Enhancing automated driving with human foresight," in *Proc. IEEE Intell. Vehicles Symp. Workshops*, Jul. 2021, pp. 25–30.
[67] K. Chitta, A. Prakash, and A. Geiger, "NEAT: Neural attention fields for end-to-end autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15773–15783.
[68] J. Dong, S. Chen, S. Zong, T. Chen, and S. Labi, "Image transformer for explainable autonomous driving system," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2732–2737.
[69] J. P. Hanna, A. Rahman, E. Fosong, F. Eiras, M. Dobre, J. Redford, S. Ramamoorthy, and S. V. Albrecht, "Interpretable goal recognition in the presence of occluded factors for autonomous vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 7044–7051.
[70] H. Mankodiya, M. S. Obaidat, R. Gupta, and S. Tanwar, "XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles," in *Proc. Int. Conf. Commun., Comput., Cybersecurity, Informat. (CCCI)*, Oct. 2021, pp. 1–5.
[71] A. S. Madhav and A. K. Tyagi, "Explainable artificial intelligence (XAI): Connecting artificial decision-making and human trust in autonomous vehicles," in *Proc. 3rd Int. Conf. Comput., Commun., Cyber-Secur.*, 2021, pp. 123–136.
[72] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering counterfactual explanations with semantics," in *Proc. 17th Eur. Conf.*, 2022, pp. 387–403.
[73] Z. Zhang, R. Tian, R. Sherony, J. Domeyer, and Z. Ding, "Attention-based interrelation modeling for explainable automated driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 2, pp. 1564–1573, Feb. 2023.
[74] S. Kolekar, S. Gite, B. Pradhan, and A. Alamri, "Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net model with grad-CAM visualization," *Sensors*, vol. 22, no. 24, p. 9677, Dec. 2022.
[75] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "OCTET: Object-aware counterfactual explanations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15062–15071.
[76] M. Itkina and M. Kochenderfer, "Interpretable self-aware neural networks for robust trajectory prediction," in *Proc. Conf. Robot Learn.*, 2023, pp. 606–617.
[77] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 1–12, Sep. 2023.
[78] H. Hu, Q. Wang, Z. Zhang, Z. Li, and Z. Gao, "Holistic transformer: A joint neural network for trajectory prediction and decision-making of autonomous vehicles," *Pattern Recognit.*, vol. 141, Sep. 2023, Art. no. 109592.

[79] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, "Explaining autonomous driving actions with visual question answering," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 1207–1214.
[80] J. Echterhoff, A. Yan, K. Han, A. Abdelraouf, R. Gupta, and J. McAuley, "Driving through the concept gridlock: Unraveling explainability bottlenecks in automated driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7346–7355.
[81] Y. Feng and Y. Sun, "PolarPoint-BEV: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 1–11, Oct. 2024.
[82] S. Arfini, P. Bellani, A. Picardi, M. Yan, F. Fossa, and G. Caruso, "Design for inclusivity in driving automation: Theoretical and practical challenges to human-machine interactions and interface design," in *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Cham, Switzerland: Springer, 2023, pp. 63–85.
[83] P. Langley, "Varieties of explainable agency," in *Proc. ICAPS Workshop Explainable AI Planning (XAIP)*, 2019, pp. 1–11.
[84] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 279–288.
[85] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[86] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2425–2452, Oct. 2022.
[87] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, Dec. 2000, pp. 241–250.
[88] T. R. Roth-Berghofer, "Explanations and case-based reasoning: Foundational issues," in *Proc. Eur. Conf. Case-Based Reasoning*, 2004, pp. 389–403.
[89] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proc. 11th Int. Conf. Ubiquitous Comput.*, Sep. 2009, pp. 195–204.
[90] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–15.
[91] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–15.
[92] F. Naujoks, S. Hergeth, K. Wiedemann, N. Schömig, and A. Keinath, "Use cases for assessing, testing, and validating the human machine interface of automated driving systems," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2018, vol. 62, no. 1, pp. 1873–1877.
[93] T. Schneider, S. Ghellal, S. Love, and A. R. S. Gerlicher, "Increasing the user experience in autonomous driving through different feedback modalities," in *Proc. 26th Int. Conf. Intell. User Interface*, Apr. 2021, pp. 7–10.
[94] A. Kuznietsov, B. Gyevnar, C. Wang, S. Peters, and S. V. Albrecht, "Explainable AI for safe and trustworthy autonomous driving: A systematic review," 2024, *arXiv:2402.10086*.
[95] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[96] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 3–19.
[97] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[98] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[99] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.
[100] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, "Interpreting the predictions of complex ML models by layer-wise relevance propagation," 2016, *arXiv:1611.08191*.
[101] 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 Commun.*, vol. 10, no. 1, pp. 1–8, Mar. 2019.
[102] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[103] H. K. B. Babiker and R. Goebel, "An introduction to deep visual explanation," in *Proc. 31st Neural Inf. Process. Syst. Conf.*, 2017, pp. 1–27.
[104] H. Babiker and R. Goebel, "Using KL-divergence to focus deep visual explanation," in *Proc. 31st Neural Inf. Process. Syst. Conf. (NIPS)*, 2017, pp. 1–29.
[105] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[106] (2024). *Public Driving Dataset*. [Online]. Available: https://github.com/commaai/research
[107] (2022). *Public Driving Dataset*. [Online]. Available: https://public.roboflow.com/object-detection/self-driving-car
[108] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.
[109] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[110] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–24.
[111] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106685.
[112] S. Milani, N. Topin, M. Veloso, and F. Fang, "Explainable reinforcement learning: A survey and comparative review," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–36, Jul. 2024.
[113] Y. Bekkemoen, "Explainable reinforcement learning (XRL): A systematic literature review and taxonomy," *Mach. Learn.*, vol. 113, no. 1, pp. 355–441, Jan. 2024.
[114] X. Pan, X. Chen, Q. Cai, J. Canny, and F. Yu, "Semantic predictive control for explainable and efficient policy learning," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3203–3209.
[115] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5068–5078, Jun. 2022.
[116] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, Nov. 2017, pp. 1–16.
[117] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst," *Robot., Sci. Syst.*, vol. 1, no. 4, pp. 1–32, 2018.
[118] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. D. Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1389–1398.
[119] L. M. Schmidt, G. Kontes, A. Plinge, and C. Mutschler, "Can you trust your autonomous car? Interpretable and verifiably safe reinforcement learning," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 171–178.
[120] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13731–13737.
[121] H. Wang, W. Wang, S. Yuan, and X. Li, "Uncovering interpretable internal states of merging tasks at highway on-ramps for autonomous driving decision-making," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 2825–2836, Oct. 2022.
[122] G. Rjoub, J. Bentahar, and O. A. Wahab, "Explainable AI-based federated deep reinforcement learning for trusted autonomous driving," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, May 2022, pp. 318–323.
[123] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "PlanT: Explainable planning transformers via object-level representations," in *Proc. 6th Annu. Conf. Robot Learn.*, 2022, pp. 1–9.

[124] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 673–683, Jan. 2023.

[125] B. Hejase, E. Yurtsever, T. Han, B. Singh, D. P. Filev, H. E. Tseng, and U. Ozguner, "Dynamic and interpretable state representation for deep reinforcement learning in automated driving," *IFAC-PapersOnLine*, vol. 55, no. 24, pp. 129–134, 2022.

[126] L. Cultrera, F. Becattini, L. Seidenari, P. Pala, and A. D. Bimbo, "Explaining autonomous driving with visual attention and end-to-end trainable region proposals," *J. Ambient Intell. Humanized Comput.*, vol. 1, pp. 1–13, Feb. 2023.

[127] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Proc. Conf. Robot Learn.*, 2023, pp. 726–737.

[128] R. Paleja, Y. Niu, A. Silva, C. Ritchie, S. Choi, and M. Gombolay, "Learning interpretable, high-performing policies for autonomous driving," *Robot. Sci. Syst.*, vol. 1, no. 3, pp. 1–47, 2022.

[129] E. M. Kenny, M. Tucker, and J. Shah, "Towards interpretable deep reinforcement learning with human-friendly prototypes," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–36.

[130] Q. Yang, H. Wang, M. Tong, W. Shi, G. Huang, and S. Song, "Leveraging reward consistency for interpretable feature discovery in reinforcement learning," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 54, no. 2, pp. 1014–1025, Feb. 2024.

[131] H. Lu, Y. Liu, M. Zhu, C. Lu, H. Yang, and Y. Wang, "Enhancing interpretability of autonomous driving via human-like cognitive maps: A case study on lane change," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 3, pp. 1–11, Dec. 2024.

[132] W. Liu, D. Li, E. Aasi, R. Tron, and C. Belta, "Interpretable generative adversarial imitation learning," 2024, *arXiv:2402.10310*.

[133] C. Wang and N. Aouf, "Explainable deep adversarial reinforcement learning approach for robust autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 1–13, Feb. 2024.

[134] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[135] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION dataset: An international, adversarial and cooperative moTION dataset in interactive driving scenarios with semantic maps," 2019, *arXiv:1910.03088*.

[136] D. Omeiza, H. Web, M. Jirotka, and L. Kunze, "Towards accountability: Providing intelligible explanations in autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 231–237.

[137] C. Brewitt, B. Gyevnar, S. Garcin, and S. V. Albrecht, "GRIT: Fast, interpretable, and verifiable goal recognition with learned decision trees for autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 1023–1030.

[138] Z. Cui, M. Li, Y. Huang, Y. Wang, and H. Chen, "An interpretation framework for autonomous vehicles decision-making via SHAP and RF," in *Proc. 6th CAA Int. Conf. Veh. Control Intell. (CVCI)*, Oct. 2022, pp. 1–7.

[139] U. Onyekpe, Y. Lu, E. Apostolopoulou, V. Palade, E. U. Eyo, and S. Kanarachos, "Explainable machine learning for autonomous vehicle positioning using SHAP," in *Explainable AI: Foundations, Methodologies and Applications*, 2023, pp. 157–183.

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

[141] J. Ayoub, N. Du, X. J. Yang, and F. Zhou, "Predicting driver takeover time in conditionally automated driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9580–9589, Jul. 2022.

[142] C. Brewitt, M. Tamborski, C. Wang, and S. V. Albrecht, "Verifiable goal recognition for autonomous driving with occlusions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 11210–11217.

[143] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger. (2024). *PlanT Project Homepage*. [Online]. Available: https://www.katrinrenz.de/plant/

[144] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[145] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6254–6263.

[146] L. De Moura and N. Bjørner, "Z3: An efficient SMT solver," in *Proc. Theory Pract. Softw., Int. Conf. Tools Algorithms Construct. Anal. Syst.*, 2008, pp. 337–340.

[147] J. Suchan, M. Bhatt, and S. Varadarajan, "Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1879–1885.

[148] A. Corso and M. J. Kochenderfer, "Interpretable safety validation for autonomous vehicles," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[149] J. DeCastro, K. Leung, N. Aréchiga, and M. Pavone, "Interpretable policies from formally-specified temporal properties," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.

[150] S. Kothawade, V. Khandelwal, K. Basu, H. Wang, and G. Gupta, "AUTO-DISCERN: Autonomous driving using common sense reasoning," 2021, *arXiv:2110.13606*.

[151] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[152] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[153] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, and H. Hussmann, "I drive–you trust: Explaining driving behavior of autonomous cars," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–6.

[154] G. Wiegand, M. Eiband, M. Haubelt, and H. Hussmann, "'I'd like an explanation for that!' Exploring reactions to unexpected autonomous driving," in *Proc. 22nd Int. Conf. Hum.-Comput. Interact. Mobile Devices Services*, Oct. 2020, pp. 1–11.

[155] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fulbier, and A. R. S. Gerlicher, "Explain yourself! Transparency for positive UX in autonomous driving," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–12.

[156] T. Schneider, J. Hois, A. Rosenstein, S. Metzl, A. R. S. Gerlicher, S. Ghellal, and S. Love, "Don't fail me! The level 5 autonomous driving information dilemma regarding transparency and user experience," in *Proc. 28th Int. Conf. Intell. User Interface*, Mar. 2023, pp. 540–552.

[157] G. Kim, D. Yeo, T. Jo, D. Rus, and S. Kim, "What and when to explain? On-road evaluation of explanations in highly automated vehicles," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 3, pp. 1–26, Sep. 2023.

[158] Y. Shen, S. Jiang, Y. Chen, and K. R. Driggs-Campbell, "To explain or not to explain: A study on the necessity of explanations for autonomous vehicles," in *Proc. NeurIPS Workshop Prog. Challenges Building Trustworthy Embodied AI*, 2022, pp. 1–38.

[159] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Tech. Rep., 2018.

[160] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 1–14.

[161] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[162] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[163] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[164] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[165] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90% ChatGPT Quality*. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[166] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following LLaMA model," Tech. Rep., 2023.

[167] *Introducing Claude*, Anthropic, 2023.

[168] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23716–23736.
[169] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–26.
[170] D. Driess et al., "PaLM-E: An embodied multimodal language model," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 8469–8488.
[171] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning united visual representation by alignment before projection," 2023, *arXiv:2311.10122*.
[172] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," in *Proc. Conf. Empirical Methods Natural Lang. Processing, Syst. Demonstrations*, 2023, pp. 543–553.
[173] G. Team et al., "Gemini: A family of highly capable multimodal models," 2023, *arXiv:2312.11805*.
[174] Anthropic. (2024). *Introducing the Next Generation of Claude*. [Online]. Available: https://www.anthropic.com/news/claude-3-family
[175] Waywe Res. (2023). *LINGO-1: Exploring Natural Language for Autonomous Driving*. [Online]. Available: https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/
[176] Waywe Res. Team. (2024). *LINGO-2: Driving With Natural Language*. [Online]. Available: https://wayve.ai/thinking/lingo-2-driving-with-language/
[177] L. Chen, O. Sinavski, J. Hunermann, A. Karnsund, A. James Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," 2023, *arXiv:2310.01957*.
[178] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," 2023, *arXiv:2310.01412*.
[179] A.-M. Marcu, L. Chen, J. Hunermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton, E. Arani, and O. Sinavski, "LingoQA: Video question answering for autonomous driving," 2023, *arXiv:2312.14115*.
[180] S. Park, M. Lee, J. Kang, H. Choi, Y. Park, J. Cho, A. Lee, and D. Kim, "VLAAD: Vision and language assistant for autonomous driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 980–987.
[181] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," in *Proc. NeurIPS Found. Models Decis. Making Workshop*, 2023, pp. 1–28.
[182] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. Eben Li, M. Tomizuka, W. Zhan, and M. Ding, "LanguageMPC: Large language models as decision makers for autonomous driving," 2023, *arXiv:2310.03026*.
[183] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "DiLu: A knowledge-driven approach to autonomous driving with large language models," in *Proc. Int. Conf. Learn. Represent.*, 2024, pp. 59–67.
[184] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "RAG-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," 2024, *arXiv:2402.10828*.
[185] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. Murthy Jatavallabhula, and K. M. Krishna, "Talk2BEV: Language-enhanced Bird's-eye view maps for autonomous driving," 2023, *arXiv:2310.02251*.
[186] S. Atakishiyev, M. Salameh, and R. Goebel, "Safety implications of explainable artificial intelligence in end-to-end autonomous driving," 2024, *arXiv:2403.12176*.
[187] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," 2023, *arXiv:2307.07162*.
[188] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving," 2023, *arXiv:2312.03661*.
[189] F. Chi, Y. Wang, P. Nasiopoulos, and V. C. M. Leung, "Multi-modal GPT-4 aided action planning and reasoning for self-driving vehicles," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 7325–7329.
[190] Y. Duan, Q. Zhang, and R. Xu, "Prompting multi-modal tokens to enhance end-to-end autonomous driving imitation learning with LLMs," 2024, *arXiv:2404.04869*.
[191] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards safe, explainable, and regulated autonomous driving," *Explainable Artificial Intelligencce Intelligent Transportation Systems*, vol. 1, pp. 32–52, Nov. 2023.
[192] *Road Vehicles—Functional Safety—Part 6: Product Development at the Software Level*, Standard ISO 26262-6, 2018. [Online]. Available: https://www.iso.org/standard/68388.html
[193] T. A. Johansen, T. Perez, and A. Cristofaro, "Ship collision avoidance and COLREGS compliance using simulation-based control behavior selection with predictive hazard assessment," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3407–3422, Dec. 2016.
[194] S. L. Hantler and J. C. King, "An introduction to proving the correctness of programs," *ACM Comput. Surveys*, vol. 8, no. 3, pp. 331–353, Sep. 1976.
[195] *Road Vehicles—Cybersecurity Engineering*, Standard ISO/SAE 21434:2021, ISO Tech. Commitee, 2021.
[196] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 998–1026, 2nd Quart., 2020.
[197] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Comput. Secur.*, vol. 103, Apr. 2021, Art. no. 102150.
[198] X. Sun, F. R. Yu, and P. Zhang, "A survey on cyber-security of connected and autonomous vehicles (CAVs)," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6240–6259, Jul. 2022.
[199] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *Int. J. Interact. Design Manuf. (IJIDeM)*, vol. 9, no. 4, pp. 269–275, Nov. 2015.
[200] J. Haspiel, N. Du, J. Meyerson, L. P. Robert, D. Tilbury, X. J. Yang, and A. K. Pradhan, "Explanations and expectations: Trust building in automated vehicles," in *Proc. Companion ACM/IEEE Int. Conf. Human-Robot Interact.*, Mar. 2018, pp. 119–120.
[201] G. Huang and B. J. Pitts, "Takeover requests for automated driving: The effects of signal direction, lead time, and modality on takeover performance," *Accident Anal. Prevention*, vol. 165, Feb. 2022, Art. no. 106534.
[202] B. Mok, M. Johns, K. J. Lee, D. Miller, D. Sirkin, P. Ive, and W. Ju, "Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2458–2464.
[203] J. Wan and C. Wu, "The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 6, pp. 582–591, Dec. 2018.
[204] A. Schieben, M. Wilbrink, C. Kettwich, R. Madigan, T. Louw, and N. Merat, "Designing the interaction of automated vehicles with other traffic participants: Design considerations based on human needs and expectations," *Cognition, Technol. Work*, vol. 21, no. 1, pp. 69–85, Feb. 2019.
[205] S. Thrun et al., "Stanley: The robot that won the DARPA grand challenge," *J. Field Robot.*, vol. 23, no. 9, pp. 661–692, 2006.
[206] A. Jain, L. Del Pero, H. Grimmett, and P. Ondruska, "Autonomy 2.0: Why is self-driving always 5 years away?" 2021, *arXiv:2107.08142*.
[207] J. Hawke, H. E, V. Badrinarayanan, and A. Kendall, "Reimagining an autonomous vehicle," 2021, *arXiv:2108.05805*.
[208] E. Dagan. (2024). *Solving the Long-Tail With E2E AI: The Revolution Will Not Be Supervised?*. [Online]. Available: https://wayve.ai/thinking/e2e-embodied-ai-solves-the-long-tail/
[209] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied AI: From simulators to research tasks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 230–244, Apr. 2022.
[210] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024.
[211] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," 2016, *arXiv:1606.06565*.
[212] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–63.
[213] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Cooperative autonomous vehicles that sympathize with human drivers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 4517–4524.

[214] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone, "Reward (Mis)design for autonomous driving," *Artif. Intell.*, vol. 316, Mar. 2023, Art. no. 103829.
[215] (2020). *The Molly Problem*. [Online]. Available: https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/MollyProblem.aspx
[216] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3145–3153.
[217] S. Magdici and M. Althoff, "Fail-safe motion planning of autonomous vehicles," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 452–458.
[218] W. Xue, Z. Wang, R. Zheng, X. Mei, B. Yang, and K. Nakano, "Fail-safe behavior and motion planning incorporating shared control for potential driver intervention," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 1–15, Oct. 2023.
[219] C. Pek and M. Althoff, "Fail-safe motion planning for online verification of autonomous vehicles using convex optimization," *IEEE Trans. Robot.*, vol. 37, no. 3, pp. 798–814, Jun. 2021.
[220] E. Kenny and J. Shah, "In pursuit of regulatable LLMs," in *Proc. NeurIPS Workshop Regulatable ML*, 2023, pp. 1–19.
[221] L. Sanneman and J. A. Shah, "The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems," *Int. J. Human–Computer Interact.*, vol. 38, nos. 18–20, pp. 1772–1788, Dec. 2022.
[222] M. R. Endsley, "Supporting human-AI teams: Transparency, explainability, and situation awareness," *Comput. Hum. Behav.*, vol. 140, Mar. 2023, Art. no. 107574.
[223] T. Zhang, W. Li, W. Huang, and L. Ma, "Critical roles of explainability in shaping perception, trust, and acceptance of autonomous vehicles," *Int. J. Ind. Ergonom.*, vol. 100, Mar. 2024, Art. no. 103568.
[224] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101896.
[225] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, and M. O. Riedl, "Operationalizing human-centered perspectives in explainable AI," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–6.

SHAHIN ATAKISHIYEV (Graduate Student Member, IEEE) received the B.Sc. degree in computer engineering from Qafqaz University, Azerbaijan, in June 2015, and the M.Sc. degree in computer engineering (software engineering and intelligent systems) from the University of Alberta, Canada, in January 2018, where he is currently pursuing the Ph.D. degree with the Explainable Artificial Intelligence (XAI) Laboratory, Department of Computing Science, under the supervision of Prof. Randy Goebel. His research interests include safe, ethical, and explainable artificial intelligence applied to real-world problems.

MOHAMMAD SALAMEH received the Ph.D. degree from the University of Alberta, with a focus on statistical machine translation and sentiment analysis, under the supervision of Dr. Greg Kondrak and Dr. Colin Cherry. He is currently a Principal Researcher with Huawei Technologies Canada Company Ltd., and leading the Neural Architecture Search Group, focusing on gradient-based and reinforcement learning approaches. He co-organized the Determining Sentiment Intensity in Tweets (SemEval2016) and the 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 the 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. Previously and while working on this paper, he was with Huawei Technologies Canada Company Ltd., Edmonton, AB, Canada. He is currently a Senior Research Scientist with Sony AI.

RANDY GOEBEL received the B.Sc. degree in computer science from the University of Regina, the M.Sc. degree in computing science from the University of Alberta, and the Ph.D. degree in computer science from The University of British Columbia.

He held faculty appointments at the University of Waterloo; The University of Tokyo; Multimedia University, Kuala Lumpur; and Hokkaido University, Sapporo. He was a Visiting Researcher with the 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 U.K, and China. He is currently a Professor of computing science with the Department of Computing Science, University of Alberta. He is also a Fellow and the Co-Founder of Alberta Machine Intelligence Institute (Amii). His theoretical work on abduction, hypothetical reasoning, and belief revision is internationally well known. He has worked on optimization, algorithm complexity, systems biology, natural language processing, and automated reasoning. His research interests include the formalization of visualization and explainable artificial intelligence (XAI), especially in applications in autonomous driving, legal reasoning, and precision health.