Explainable AI for Transparent Decision-Making in Autonomous Vehicle Navigation

By Dr. Bart Van Daele

Professor of Electrical Engineering, Ghent University, Belgium

1. Introduction

[1] A key concept for navigational algorithms for self-driving vehicles is the necessity to be explainable. This leads to the requirement for transparent algorithms. Understandably, this requirement describes the ability of a system to understand the underlying principles involved in the navigational decisions being made. This article clearly demonstrates the importance of these requirements in the context of future architectures for assisted and autonomous driving. The proposed architecture completely decouples control law from an auxiliary system model, thus providing explainability via the system's transfer functions. In the literature, there are a few of works which study such issues.[2] There is increasing concern over the operation of the modern society, with an increasing amount of human decisions being replaced by artificially intelligent systems. This has led to an increasing body of research that aims to understand and implement transparency into decision-making systems by mechanisms such as interpretable deep learning. It is critical that we develop the capabilities to explain decisions made by systems that can significantly impact human life. In this work we have attempted to present recent advances in transparent and explainable AI and to expose potential physical explanations for similarities and differences between decisional processes made by human drivers and those made by 'autonomous' machines that operate in similar conditions.

1.1. Background and Motivation

Explainable AI (XAI) systems need to answer a user's "why" question in an understandable way [3]. The aim of the SHADE project is to improve the accuracy of the AI based Autonomous Vehicle (AV) system, using the explainability of AI as a key factor. We do this by engaging real users in a user-centred design process, identifying user requirements, and evaluating how explanations can influence the drivers' trust and transparency. An EU-study

found that 248 out of 255 participants felt AI-driven AVs should always be equipped with an explainability system.

Autonomous vehicles are predicted to eliminate 94% of vehicle crashes, significantly reduce traffic commute times, and make mobility more accessible [4]. Notwithstanding these advantages, safety and transparency (or the lack thereof) in AI systems have been important contributing factors to low public-trust in autonomous vehicles. Explainability can also potentially lead to better Human-AI interfaces because it can enable humans to identify issues in the AI system a priori without the presence of artificial adversity [5].

1.2. Scope and Objectives

In the daily interactions between humans and AI, human users can only really trust an AI application if they understand, or at least have awareness of, the thought-process behind its decisions. AI systems are designed and created by humans, and this academic paper "Towards Safe, Explainable, and Regulated Autonomous Driving" shows that the deployment of autonomous cars is projected to confer large benefits by, for example, significantly reducing travel time and the number of road traffic fatalities and severe injuries. In fact, the OECD estimates suggest that, if all motor vehicles were to become fully autonomous, travel times might decrease by at least 10 % and the number of fatalities and severe injuries might be reduced by 60 % to 70 % [4].

Explainable AI has gained a remarkable amount of importance across the AI community. According to Thomas Arnold, Director in Artificial Intelligence at Improbable, rule-based AI is usually more transparent than the newer, algorithmic models. This made BUILD Engineering wonder: could more traditional models for navigation in a computer game be used in a similar way for an autonomous vehicle? In addition, biases in AI systems can be seriously harmful when they are applied in safety-critical situations or in morally sensitive contexts. Model interpretability makes it possible to understand and validate a model and makes it easier to monitor the outputs, consider edge cases (e.g. for corner cases, the algorithm should inform the operator of the or the input data), and thus address it's potential biases. The use of nonparametric modeling is more intuitive among physiologists and engineers and more easily interpreted than the use of a complicated off-the-shelf deep model in most cases [6].

2. Fundamentals of Autonomous Vehicle Navigation

Interpretable issues and ethical concerns in unmanned driving hinder its advancement. Explaining self-driving behavior is essential for safety-critical applications and to enhance user trust. The wide application of neural networks has sparked interest in their transparency and interpretability [2]. The internal mechanisms of deep neural networks, such as synaptic weights and the interpretation of data, typically remain unknown to human operators, which can lead to disastrous errors in systems that rely upon them. Decrypting the intermediate process of deep neural networks can improve their credibility and acceptability. In autonomous vehicles, the reliability and interpretability of deep neural network outputs are crucial, especially in addressing ethical issues. Autonomous vehicles often use deep multi-modal learning and object detection techniques to anticipate and avoid collisions. The learned navigation model is transferrable to real and unknown settings but may have poor performance in open spaces with reflective surfaces, due to drastic changes in optical flow representations encountered when transferring learned knowledge to true data. Retraining navigation models downstream using true data has been shown to enhance collision avoidance in such reflective spaces. All of these AI-based navigation technologies are typically trained offline, which makes them inapplicable to critical driving situations. An approach that fuses classical model predictive control with learned models has recently been introduced.

Safely navigating in autonomous vehicles requires an accurate representation of the environment and the vehicle's positioning. Traditional methods like GPS and RTK have limitations, especially in complex urban or tunnel scenarios [7]. Simultaneous Localization and Mapping (SLAM) is a promising solution for autonomous vehicle positioning and navigation, with LIDAR SLAM and Visual SLAM being the two main categories. In visual SLAM, which doesn't require any infrastructure, traditional methods use feature-based detection algorithms, while more recent methods based on deep learning have shown consistent performances. Deep learning has also been applied in LIDAR SLAM for the feature representation and system optimization. Real-world examples of AI and autonomous driving synergy include Tesla Autopilot, Waymo, and Mobileye. SLAM and AI can also be combined to generate high-definition maps for autonomous vehicle navigation and enable behavior prediction for autonomous vehicles, based on the clustering and classification of specific lane-change path features (methods improve both model accuracy and execution speed) [8].

2.1. Basic Concepts and Components

In this context, the solution and navigational processes of the deep learning model developed for autonomous vehicle navigation are explained via change visualizations, and the results obtained may be returned when necessary. It is important to verify all decision-making processes with human drivers in order to adapt this process to the traffic. Thanks to the proposed change visualizations, we will continue to inform the vehicle driver about the decision taking basis of the autonomous vehicle. In other words, one of the aims of this work, which proposes model-specific XAI methods to be used in the change visualization process, is to ensure that the level of trust in autonomous vehicles is increased and that a transparent communication is established at the decision-making stage of vehicle customers. In this sense, transparency and trust in the decision systems of autonomous vehicles play an important role in the safety and trustability of the vehicle, which constitutes an important competitive area on the road, and the determination of a new feature by human drivers in the vehicle selection will affect the corporate behavior of the vehicle lie.

The transparent interpretation of decisions made by a learning machine plays an important role in many social life concerns ranging from the assignment of bank loans to autonomous vehicle traffic. In addition to the fact that learning algorithms can now reach the performance level of human experts in many fields, the lack of transparency in learning machines also restricts their use in various areas [9]. The lack of transparency in sophisticated models such as convolutional neural networks (CNN) gained in prominence as a major obstacle. XAI (eXplainable Artificial Intelligence) methods aim to help humans and learning machines act as good partners by explaining all decision-making processes, suggesting improvements to models for decisions. The main purpose of this chapter is to adapt model-specific XAI methods into the autonomous vehicle navigation process and to ensure that the visual interpretations obtained can be used in traffic, safety, and safety communication systems at the vehicle decision-making unit level, which is the main process in the autonomous vehicle navigation process [10].

3. Explainable AI Techniques

In the next level of understanding for the vision illustrated in figures 4 and 5, detailed description of the AI models and their decisions involved, go beyond both the abstarct explanations of safety alone such as kinematically explainable regions from Section 3.2 and

also the one abstarct explantions that go with the AI's own understanding of the environment that is given by the set of the local viewpoint obtained close to the time of decision making [11]. This is illustrated here via the abstract graphs of the model which can visualize the internal model transitions that facilitate decision-making at different instances inside the model. This granularty of understanding will be able to improve safety features by diagnosing near-aptic failure conditions at a granular level such as those pertaining to dynamics, and model level beliefs and actions. At the same time, AI is responsible for long-term commitment of the overall objectives and the social needs that create need for high-level acounterabilities.

The overall vision of full-fledge AI decision-making still involves a step-wise approach where AI can potentially support the human driver or the human in the loop in a responsible and explainable manner. This final aim has already resulted in appropriate but separate perspectives regarding AIT and AV navigation, where explanations fulfill distinct requirements of human involvement and intent as pointed out earlier in Section 2. For AIT one focuses on model explaining to backup AI trust without immediate impacts on task outcome. This covers an interim stage for safe navigation illustrated here for an electric scooter model in the illustrative city scenario. In the AV navigation context in contrast, one needs explanations to account what happened during accident or trouble spots correctly and this calls for explanations with explainable units of AV understanding/sensing as explained via our local navigation strategy over the simplistic grid scenario as explained in Figure 1.

3.1. Interpretable Machine Learning

Model agnostics methods have a very broad approach as they try to generate a common explanation mechanism that can explain the decision of any network regardless of its network architecture. When applied to deep learning models and images, these explanations do not give any clear or as such interpretable explanation of what features the model is looking at. Importance of Global and Local methods can be understood from the fact that it is these local and global structural characteristics, which are important in understanding the physical significance of the relevant features in DNN, that vary on a model level and thus are of more interest. Thus, it is clear that the need of the hour is to indeed develop a structural explanation mechanism such that, the explanations generated retains the global and local characteristic of the model, this is referred to as structured explanations [12].

Comprehensibility in overall decision-making abilities of Artificial Intelligence (AI) systems is crucial not only for improving users' confidence but also for ensuring the system accordingly taking care of end-users' responses and suggestions. Explainable AI (XAI) is an essential requirement to show why a particular decision is being made in a particular case. In [13] a new AI episode in the teenage of the 21st century was initiated with robust new algorithms such as deep learning and the episode is continuing with the development of the highly complicated modern AI network such as ResNet, GPT-3, and many others. In AI development, computation has been establishing itself as an alternative to understanding and humans now believe that such a novel learning dynamics provides superior decision-making ability in front of complex data streams. The mathematical principles behind these complicated learning architectures are not well resolved, thus, it is indeed challenging to explain AI decision-making policies beyond the accuracy level, fraction of correct decisions, as the ultimate metric. Therefore, it is believed that the concept of XAI will be the main agent to future-proof AI learning solutions. XAI has been gaining much momentum in the AI development domain as it focuses on verifiable models, interpretable results, transparent working principles as well as direct alignments with explainability-and-accountability compliance. XAI serves as a problem-driven corrective domain to AI with the main goal to expose the logical-justification and decision-making process for any given input observation.

3.2. Model Explanation Techniques

Highly interpretable models such as decision trees have been formulated to prioritize feature interpretability and thereby, model performance. On the contrary, well-established highly performing models require interpretable explanations [6]. Class Activation Mapping (CAM) and saliency maps are popular techniques in computer vision which provide visual evidence as to what in the feature space is contributing to model classification decisions. Grad-CAM and Integrated Gradient better the visual evidence provided by localization. These exist to both trace and interpret the features in the classifier, and to understand the features that drive the AI model decisions. Within autonomous vehicles, visual spatial explanation not only improves human trust in the system, but also reduces cognitive load in making a decision about correcting the poor AI choice. It is therefore crucial in society that autonomous vehicles are governed by a human-in-the-loop, in partnership with human trust that is engendered by model explanations.

AI models are powerful tools that can optimize decision-making at a level humans are incapable of. Often, however, the reason behind the decisions made by these models is unknown, leading to public lack of trust, and, more importantly, lack of understanding from specialists in the field. This can be disastrous in ethical and moral fields such as medicine or law. For example [14], AI models predicted mental illness in Singaporeans with 88% accuracy but without providing effective explanation or reasoning behind them. To address this challenge, model explanation or interpretability has to be demonstrated for decision-making in high-stakes industries such as medicine. Explainable AI (XAI) is an active and growing area within AI, aiming to address the concerns of bias, trust, understanding, and ethical issues in societal decision-making.

4. Applications of Explainable AI in Autonomous Vehicle Navigation

[9] Explainable AI is an essential component in improving algorithmic trust and compliance in the autonomous vehicle sector. IXAI (Interactive Explainable AI) is used to detect and classify objects around autonomous vehicles, with a focus on delivering uncertainty and abnormality at reasoning time. A new feature explanation framework is proposed for the autonomous vehicle system, encompassing fine-grained, object-centric feature explanations. Advanced driver-assistance systems (ADAS) form the basis of autonomous vehicle navigation, various object-detection systems perform well in road-scene navigation and a driving simulator typically encompasses all the navigation tasks that exist in real-world road systems. The proposed bound tensor decomposition method is applied for creating reusable and explainable content for a more diverse set of vision tasks and autonomous vehicle navigation [15] A detailed study on the application of Reinforcement Learning (RL) in controlling autonomous vehicles deployed in modern road systems, the core of which is LiDAR + Camera motor sensors and visuo-motor-predictive systems in foveated vision. The implemented system reaches a high score (90%) at a familiar environment and also has a generalization to unseen road types in any simulated urban environment in terms of human-machine interfaces. In the present rally, the need for introducing human-like rules to interpret navigation commands and to program visually explainable behaviors in abstract situations that appear during road scene navigation is shown.[4] This virtual test bench contains detailed scenario descriptions to be executed virtually in physical robot drive responding to real-time sensor data. The real-time performance includes pedestrian, vehicle, and object detection tasks from camera, 3D bounding box detection from LiDAR, and affordance concept detection

using the object properties of detected objects in the autonomous vehicle navigation decisions. Besides creating a virtual test track to navigate, this benchmark introduces plausible daily life scenarios in a traffic scene while point-based scalar affirmations annotated from the end-users are also included. Participants can train their models and algorithms on these scenarios to make sure their system is plausible and compliant with end-user affirmations. For efficient navigation, new interactive benchmark Sparrow-AI, the interaction between human and autonomous system, eye-tracking, generator adversarial network environment navigation, image vocabulary navigation, long intersection networking tangram navigation, predicting intersection post-occupancy arrival time, and autonomous vehicle-pedestrian spatial interaction. Autonomous vehicles (AVs) which are capable of driving on public roads, applicable to a wide variety of new roads and able to copy from human-like behavior in navigation are rapidly entering into the daily life of the public raising a concern for the explanation of why AVs act. In this context, it is still difficult to make a firm position on how driving strategies are generated with deep neural networks (DNNs), especially in complex and confusing traffic scenarios. We still focus on turning the black-box driving behavior of DNNs into white-box models. Most of the contributions are obtained also from the developed forward and inverse models in AV navigation but we still suffer from over-states prediction accuracy even for the same state. The developed modular software is capable of learning a wide variety of new roads and also has a search space of 32^8 possibilities for the navigational symbol selection at the intersection. If we change the programmable parameters at the intersections in the simulation, the real-time LSTM system gives the capability to recurrently update the navigation and decision strategy during continuous autonomous navigation.

4.1. Route Planning and Optimization

The route-planning and optimization module plays a pivotal role in computing collision-free and efficient paths from the ego vehicle to a set point-defined or continuous goal control set within the world coordinate frame, containing obstacles and traffic rules. A variety of graph-based planning strategies, such as dynamic programming, A* search, potential field methods, sampling, and heuristic approaches, are well-represented in the literature. In this paper, we focus on one such demonstration to present an explainable AI in the route-planning/optimization module. This is specifically a combination of incrementally generated cost maps for evolving robot systems with debugging AI algorithms via novel, interpretable, and less error-prone path-iteration problems. The critical design features include a memory-

less path approximation, hierarchical multi-fidelity simulations, macro completions, collision resolutions, edge orderings, and cost map updates, based on planned-edge completions [14].

4.1. Route Planning and Optimization

Autonomous vehicle (AV) navigation is modeled as a modular system [15]. In the sense-plan-act architecture, the decision-making process can be broken down into three key modules: localization, route planning, and control [16]. Each of these components requires a certain degree of explainability to ensure the implementations of safe planning. This section of the paper looks closely into the explainable artificial intelligence (XAI) pertinent to the route-planning and optimization module from the sense-plan-act loop.

4.2. Collision Avoidance Systems

A Dual-loop Bio-inspired Controller Model developed by the 'Department of Mechanical and Mechatronics Engineering, University of Waterloo' models approaches in which the generated decision is deterministic. The quickest dangerous situation is one in which inertial forces are equivalent in magnitude to the external forces directly within the last approaching distance of 15 meters. Certain situations involve the forced entering of the central zone formed around the interacting go behaviors that were considered safe initially and accelerated. A notion is that selection of action by rational control does allow the accurate foreseeing of specified outcomes induced by environmental patterns but it does not have the ability to understand the natural russing and initiating behaviors of the individuals. It should also be noted that a displaced and flick behavior are arms of particular interest from 3D modeling views to exploit Dead Zone Avoidance Mechanism's approach. Furthermore, during the modeling of ASCA S's current decision, the target and unless opds do decide combinedly that D outputs are bad and therefore they determine to distract the D outputs without any further second thought to pursue the current rapidly changing conditions [4].

The 'Center for Dentistry' uses the collision avoidance case study to explore Explainable AI in autonomous driving [5]. In it, a target vehicle moving to the right starts driving towards the ego vehicle at point x t seconds. The vehicle moving to the right reaches a distance of d_t meters from the ego vehicle over the time interval t after the decision is made from d_0= 26.380 meters at t= 0 to d_t= 10.890 meters at t= 3.2033 seconds. From the distribution of distance moments of the hazardous configurations at t= 3.2033 seconds, it is found that an action to

slight right turn is made at t= 0.05 seconds with 0.5 probability, and a decision to hard right turn is made at t=2 seconds with 0.45 probability [14]. The threshold t_t = 3.2033 seconds for action to hard right turn reveals the certainty of safe execution. Traffic Collaborative Adaptive Cruise Control is developing approaches for intelligent autonomous vehicles to optimally navigate through dynamic flow distribution sets while waiting to ensure safe merging into traffic conditions. Rather than using rigid algorithms or human designed cost functions, the auto-tuning and update law modulates the choice of flow distribution set such as to reduce the defined objective function. In doing so, the vehicle will be able to safely merge onto highways and accelerate into decelerating traffic correctly.

5. Challenges and Limitations

Developers should focus on developing an autonomous navigation system with no central coordination while avoiding unexpected emergent-destructive phenomena and cyclomatic complexity within the system. Often, multiple sub-components of the navigation system are designed and developed asynchronously. This creates a potential for difficult-to-analyze emergent effects that may result in unexpected system behavior. Additionally, it is important to consider cycling the interaction of our navigation agents with other agents and entities in the environment [15]. Under a simplistic model, this property would not only encourage a safety-first attitude but would also prevent the learning of any aggressive/ careless strategies. A real-world navigation problem, however, requires recursively considering the possible interactions and reactions of each agent within a wider environment given any particular initial strategy [14].

The development of a robust AI system by its nature is a non-trivial undertaking due to the complexity entailed by the interaction of the system's multiple components [2]. Researchers must address and manage various challenges and limitations as they look to develop a reliable and transparently-able AI (XAI) system for autonomous navigation. Limitations are most often elusive and arise as a result of novel and complex difficulties when designing a system. The worst-case scenario for developing autonomous navigation systems involves facing an emergent-destructive phenomena that are difficult to predict, plan, and account for.

5.1. Data Privacy and Security Concerns

The realization of high-level autonomous driving has attracted worldwide interest and investment from different industries and entities. In recent years, self-driving technology based on artificial intelligence (AI), especially deep learning technology, has made rapid progress. A large amount of data processing is the core link of artificial intelligence technology. These data collect hundreds of thousands of kilometers of real vehicle user driving data from different vehicle sensors and are used for ground tracking, vehicle control, and vehicle to vehicle or vehicle to infrastructure existing and new AI technologies. AI design in autonomous vehicles must be transparent and explainable. Key questions include the type of decision matrix algorithm used, the number of decision models involved, and how their decisions are combined. Additionally, explanations for vehicle motion dynamics and factors affecting deceleration are crucial. Not much work has been done in this area, but some others have proposed explainable mechanisms for autonomous vehicle navigation [1].

Because of the high level of autonomy in the new L3–L5 AVs and frequent data exchanges between the edge network AI and the centralized AI during real vehicle user driving scenarios, data privacy and security protection concerns have attracted much attention. With the requirement for highly automated driving, there will be significant improvement in the design of artificial intelligence technology through public roads. Some of these technologies that focus on safety will certainly enhance active safety to prevent accidents. However, overdependence on artificial intelligence design inevitably brings a string of challenges that must be addressed before implementation. These are largely responsibility and transparency. Because of the high level of autonomy in such new L3–L5 automated vehicles, and the frequent data exchanges between the edge network AI and the centralized AI during real vehicle user driving scenarios, data privacy and security protection concerns have attracted much attention. These topics are highly interdisciplinary and fast moving, and so raise fresh challenges and bring previously distinct research communities together.

6. Future Directions and Research Opportunities

The processed models include Deep Reinforcement Learning (DRL), Imitation Learning (IL), and three types of Linear Classifiers' (LC) models. Our results strongly emphasize that XAI explains models' reasoning power specifically in exaggeratedly disastrous, erratic, and unusual states, conditionalities, and contingencies. [17] Consecutively, we provide a review paper enunciating the real-time, in-the-loop, and post-hoc explanation strategies under

Guidance and Assistance sub-classes of (XAI) considerably in the context of autonomous decision-making in the simulation environment. Furthermore, the prototyped pipeline for transparent decision-making catering to autonomous vehicle navigation would contribute to the actual benchmarking and the next-generation control software taxonomy. Furthermore, the proposed taxonomy would aid in solving general cases in modular autonomous vehicles, which are still unsolved computational problems.

This article integrates the manually validated XAI methods into a pipeline for autonomous decision-making in the context of autonomous vehicle navigation. The prototyped pipeline demonstrates—an infrastructure for handling a Map Editor Tool, the Taxi-to-AI bridge catering to various deep learning and interpretable models, the analysis of actions triggered by the fusion of models, and a User Interface Broadcasting method in Gazebo. [6]

6.1. Enhancing the Interpretability of AI Models

Localization and identification of attention areas of the drivers have always been a research hotspot in recent years. Research on the image recognition of drivers' attention area includes a wide area of research, but also has the problem of model transparency and interpretability. It is necessary to explain self-driving behavior due to the high-stakes and safety-critical nature of autonomous driving. As a result, there is a need for transparency and interpretability of deep neural networks, which can significantly enhance user trust in autonomous driving systems. Decrypting the intermediate process of deep neural networks and seeking the reliability of their output improves the credibility and acceptance of the technology (Li et al., 2021). Herse et al. leveraged layerwise relevance propagation method for interpretable and transparent decision-making process in autonomous vehicle navigation. They conducted a comprehensive qualitative evaluation of premiums and benefits concerning explainability and performance, and compared it to saliency maps. Their findings endorse the use of LRP as a toolbox for the analysis of deep neural networks in the development of AI-based solutions for autonomous vehicle navigation (Herse et al., 2018).

[6] Explainable AI, including interpretable models, prioritizes feature interpretability and strong performance. [2] This directly aims to make the decision-making process transparent by ensuring that the features used by the model are traceable and understandable. Techniques like decision trees and methods in computer vision, such as class activation mapping and Grad-CAM, help in achieving this transparency. Iglovikov et al. proposed a generic post-

processing method that exploits the inherently parallel nature of artificial neural networks to create a starkly differentiable post-processing network. The post-processing network learns mapping function local to the test input, and therefore retains the learned post-processing approach rather than using the same fixed function across all inputs. The method allows us to analyze the learned behavior of the original network during classification as well as make interpretations on how to minimize biases (Iglovikov et al., 2018).

7. Conclusion and Key Takeaways

Many recent publications reflect an international trend of exploring road safety benefits reaped by applying XAI in AVs, while some others have proposed a few perspectives on AV-related ethical concerns. To the best of our knowledge, our paper is the first to address the current trends of XAI research in AVs and to draw a comprehensive overview of the coming development milestones for publicly acceptable AV autonomy. To capture the general idea of our XAI guideline and sub-sections, an extended version of the paper appears in an early preprint of this paper. Our open dataset and its illustrative use cases comport well with the diversified range of XAI algorithms implemented in a wide range of domains, and offer newcomers in the field of XAI the results of the performance of 11 state-of-the-art (SOTA) XAI methods on our dataset. Future work will involve developing more applications trained on this dataset with some other SOTA XAI models for different types of explanation generation and validation, Formulating a set of accurate prediction tasks or a general purpose SOTA XAI model that passes an authenticity test from human participants. The development of a XAI dataset derived from our proposed XAI guideline can be seen as a phase in the development of a XAI recommendation toward extracting domain-specific driving knowledges from both end-users and industry participants.

[18] [1] [2]The high number of fatalities caused by traffic accidents make road safety a major concern around the world. The development of self-drive capability for vehicles is seen as a promising way to reduce the occurrences of such accidents. The safety of an autonomous vehicle (AV) is particularly improved when it can reason and act on a large and varied set of driving situations with limited control actions. To attain regulatory-compliant operational safety, AVs have to explain their decisions, both in real-time driving scenarios and offline during software verification. Ensuring the safe operation in difficult or novel conditions that are not covered by training data, AVs should consult their human operator in advance. This

is not always feasible, and in cases when it is possible, the explanation time should be kept to a minimum. Real-time explanations are needed to allow the operator to interpret model behavior and perform any necessary intervention. Our paper provides a comprehensive overview of Explainable Artificial Intelligence (XAI) approaches for explaining AV behaviors and introducing transparency on publicly acceptable AV autonomy at both offline and online levels. For generating recommendations such as performing prediction from the relocation of the eye to a center camera in an AV, we further propose the first XAI dataset OpenAI-IAP.

8. References

1. [1] F. Hussain, R. Hussain, and E. Hossain, "Explainable Artificial Intelligence (XAI): An Engineering Perspective," 2021. [PDF]
2. [2] S. Zhao, Y. Li, J. Ma, Z. Xing et al., "Research on imaging method of driver's attention area based on deep neural network," 2022. ncbi.nlm.nih.gov
3. [3] A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa et al., "IEEE P7001: A Proposed Standard on Transparency," 2021. ncbi.nlm.nih.gov
4. [4] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards Safe, Explainable, and Regulated Autonomous Driving," 2021. [PDF]
5. [5] S. Atakishiyev, M. Salameh, and R. Goebel, "Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving," 2024. [PDF]
6. Tatineni, Sumanth. "Deep Learning for Natural Language Processing in Low-Resource Languages." *International Journal of Advanced Research in Engineering and Technology (IJARET)* 11.5 (2020): 1301-1311.
7. Vemoori, Vamsi. "Comparative Assessment of Technological Advancements in Autonomous Vehicles, Electric Vehicles, and Hybrid Vehicles vis-à-vis Manual Vehicles: A Multi-Criteria Analysis Considering Environmental Sustainability, Economic Feasibility, and Regulatory Frameworks." *Journal of Artificial Intelligence Research* 1.1 (2021): 66-98.
8. Mahammad Shaik. “Reimagining Digital Identity: A Comparative Analysis of Advanced Identity Access Management (IAM) Frameworks Leveraging Blockchain Technology for Enhanced Security, Decentralized Authentication, and Trust-Centric Ecosystems”. Distributed Learning and Broad Applications in Scientific Research, vol. 4, June 2018, pp. 1-22, https://dlabi.org/index.php/journal/article/view/2.

9. Tatineni, Sumanth. "Enhancing Fraud Detection in Financial Transactions using Machine Learning and Blockchain." *International Journal of Information Technology and Management Information Systems (IJITMIS)* 11.1 (2020): 8-15.
10. [10] H. S. Kim and I. Joe, "An XAI method for convolutional neural networks in self-driving cars," 2022. ncbi.nlm.nih.gov
11. [11] A. Halilovic and S. Krivic, "Understanding Path Planning Explanations," 2023. [PDF]
12. [12] R. Kashefi, L. Barekatain, M. Sabokrou, and F. Aghaeipoor, "Explainability of Vision Transformers: A Comprehensive Review and New Perspectives," 2023. [PDF]
13. [13] S. Ellen Haupt, W. Chapman, S. V. Adams, C. Kirkwood et al., "Towards implementing artificial intelligence post-processing in weather and climate: proposed actions from the Oxford 2019 workshop," 2021. ncbi.nlm.nih.gov
14. [14] A. John Karran, T. Demazure, A. Hudon, S. Senecal et al., "Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions," 2022. ncbi.nlm.nih.gov
15. [15] H. Zheng, Z. Zang, S. Yang, and R. Mangharam, "Towards Explainability in Modular Autonomous Vehicle Software," 2022. [PDF]
16. [16] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in Autonomous Driving: A Survey," 2021. [PDF]
17. [17] Y. Guan, H. Liao, Z. Li, G. Zhang et al., "World Models for Autonomous Driving: An Initial Survey," 2024. [PDF]
18. [18] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," 2021. [PDF]