Explainable AI Techniques for Transparency in Autonomous Vehicle Decision-Making

By Dr. Ying Liu

Associate Professor of Computer Science, Nanyang Technological University (NTU), Singapore

1. Introduction

In this chapter, we thoroughly examine and meticulously evaluate the diverse range of innovative methods that have been extensively discussed and analyzed in prominent recent studies. Our primary focus revolves around the noble objective of enhancing transparency in the context of automated vehicles, with a particular emphasis on cutting-edge vision-based systems. By delving into the depths of this captivating subject matter, we aim to unravel the intricacies and complexities associated with this fascinating field of research, elucidating the various nuances and subtleties that have emerged on the forefront of technological advancements. Through a meticulous exploration of these methods, we aspire to contribute towards the further development and progression of transparency in automated vehicles, paving the way for a future that is not only safer but also more intelligently interconnected.

Autonomous vehicles (AVs) decision making abilities are governed by machine learning models, which are black boxes characterized by the lack of transparency in their decision-making process. However, transparency in the decision-making approach is of paramount importance to promote user confidence, especially when dealing with vehicles performing critical and safety-related tasks. Techniques, also known as explainable AI, have been proposed to help end-users better understand why AI systems make a specific choice, hence promoting improved engagement and trust.

1.1. Background and Significance

This paper provides a systematic literature review on methods and techniques used to increase the transparency in the decision-making process of the current AV models. Methods found in the literature that were adopted on earlier similar models are categorized into ten layers that form a 3D cube. The layers used to organize the taxonomy are defined based on

This work is licensed under CC BY-NC-SA 4.0.

the AV requirements related to transparency in their decision-making models. Data used to build the taxonomy comprises articles published and indexed in scientific digital libraries. The review uncovers the existing scarcity in methods and techniques developed in the AV field specifically designed to enhance the transparency of decision-making models of deep neural networks that are often used in the AV implementation. This paper provides a reference for future research, and the taxonomy can be potentially extended over time as new methods and techniques are developed.

Autonomous driving systems have the potential to significantly reduce motor-vehicle accidents and fatalities associated with human errors. The development of autonomous vehicles (AVs) is a complex task that involves systems engineering and creating ethical decision-making models that will define the behavior of AVs when unforeseen incidents, with potential risk to human lives, occur. Current AVs use deep learning and neural network techniques, whose models have the limitation of being complex and 'black boxes', i.e. internal invisible models. The literature increases its emphasis on the need for making AVs reliable and their decision-making process transparent before more lives can be entrusted to the self-driving vehicles reaching the public market. Transparency represents a powerful tool to help potential users and regulatory authorities evaluate and trust the AV solution regarding human lives' safety.

1.2. Research Objectives

Autonomy in self-driving cars focuses on an integrated approach. Traditionally, to have a reliable autonomous vehicle, understanding how systems are currently solving specific issues is required. This means that algorithms must be studied and their outcomes should be understood in order to be trusted, considered safe, and even explainable for human interpretability. But in deep learning models in which network parameters are learnt from data without many human interactions, there are simplified representations of the phenomena that are modelled. Most of the recent approaches plan to combine the interpretability of decision trees with the prediction capability of deep learning models by adding an explanation system on the top of the real system. This approach uses a decision-making process through a tree structure, ensuring human interpretable predictions. These explanations are computed in a way that human users could verify that the system makes decisions like humans, but the decision-making problem is divided into stages, making

This work is licensed under CC BY-NC-SA 4.0.

decisions in different moments. This new approach, even with this very little approach, contributes to greater model transparency ensuring high accuracy.

2. Fundamentals of Autonomous Vehicles

Human driver: Zero autonomy. The full-time performance by a human being of all aspects of the dynamic driving task (DDT) required to operate a vehicle when conditional automation is not engaged. Supervised driving automation: One or more driving automation systems feature that requires a human driver to be available to take, or to be called upon to take over, specific DDT while the vehicle performs all other aspects of the DDT.

To clarify the terminology, it is important to recognize the diversity of autonomous vehicles, as different disciplines employ different terms. The automotive industry refers to vehicles as "autonomous," "assisted," or "safety-critical." At the U.S. Federal level, levels of autonomy are published in the American Society of Automotive Engineering (SAE) standards J3016, J3018, and J3136. At the European level, levels of driving automation are described in the United Nations Economic Commission for Europe (UNECE) "ADAS and automation" Working Group (WG) and United Nations World Forum for Harmonization of Vehicle Regulations (WP29) Working Party on General Safety Provisions. Common synonyms for "autonomous vehicles" are self-driving cars, driverless cars, robotic cars, and self-driving vehicles.

2.1. Definition and Types of Autonomous Vehicles

According to the classification of the Society of Automotive Engineers (SAE), autonomous vehicles can be divided into five levels: Level 0 (no automation), Level 1 (driver assistance), Level 2 (partial automation), Level 3 (conditional automation), and Level 4 (high automation). Level 0 is a vehicle in which most or all driving tasks are performed by a human driver. Level 1 is an advanced driver-assisted system that assists the driver with steering or speed control independent of the driver, but not both at the same time. Level 2 is a driving system that can simultaneously assist the driver with steering and speed control. Level 3 is an automated driving system that performs all driving functions and responds to requests to intervene, but only under certain conditions. Level 4 is an automated driving system where the vehicle can operate under predetermined conditions in an unmanned setting.

Autonomous vehicles, also known as self-driving or driverless cars, refer to vehicles equipped with an autopilot system and a variety of advanced control technologies that use artificial

This work is licensed under CC BY-NC-SA 4.0.

intelligence, sensors, and mobile communication. This makes the vehicle less reliant on manual driving and able to perform driving tasks independently in an unmanned environment. In addition to traditional traffic functions, other features provide a better driving experience, improve fuel economy, and help increase the driving safety of both road users and pedestrians. It's a new generation of vehicles that play an important role in the future.

2.2. Key Components and Technologies

However, expert knowledge in what is being learned, transparency, and clear policies are critical elements for adopting AI systems in safety-critical applications. Manufacturers need to provide clear insight into how their AI systems work, principles that guide their AI behaviors, and a clear understanding of how they make decisions. By transparently sharing enough information about the operation, developers and end-users can better understand why AI systems respond in certain ways, assess the comparative risks of different choices, and make safety-conscious decisions. In the field of autonomous vehicles, and in the broader context of AI and robotics, explainable AI (XAI) has received significant attention as a means to offer such insight.

The key components and technologies in autonomous vehicles consist of different layers including perception, computer vision, natural language processing, map building, path planning, decision-making, and control, as shown in Figure 4. There exist various techniques in each layer. For instance, many perception, computer vision, and sensor technologies such as LiDAR (Light Detection and Ranging) and RADAR (Radio Detection and Ranging) technologies contribute mainly to the first layer of perception. Such techniques enable autonomous vehicles to make well-informed decisions, following the layers of path planning, decision-making, and control. There are multiple methods for these decision-making layers including machine learning techniques, deep reinforcement learning models, or formal methods. In particular, deep neural networks have become the predominant choice for machine learning, computer vision, and control systems.

3. Explainable AI in Autonomous Vehicles

We start our presentation with a short literature review of explainable AI, returning to Autonomous Vehicles, and highlight our contributions. Next, our proposed methodology is

This work is licensed under CC BY-NC-SA 4.0.

described. Finally, we set an experimental evaluation along with some concluding remarks and future work to be done.

The main motivation behind explainable machine learning goes beyond creating a trustworthy system. Explainable models can improve user understanding, provide an early warning system regarding model behavior deviations, and help in troubleshooting model-related issues. Humans have proven to possess a higher degree of trust and confidence when dealing with a system they deeply understand. Also, fully explainable models provide increased control handling since users can override the decisions made by the machine. Explainability is a mandatory requirement in regulated domains.

3.1. Importance of Transparency in AI Decision-Making

The advent of artificial intelligence (AI) has led to autonomy in decision making for control systems. Self-driving cars using AI have taken on the challenging task of transforming transportation. As AI systems become more intelligent, AI-generated decisions may lack transparency, adding possible risks such as collusion, distortion or bias, and allowing black-box decision-making. AI explainability has been enhanced by several XAI techniques. While AI may leverage compressed data, human interpretable information is needed when decisions are rendered. A committee of explanations increases reliability and provides human-understandable reasoning for decisions. Our research focuses on an ensemble of XAI techniques to increase transparency and convey information needed to understand imminent path, car state, and feasibility predictions such as acceleration and velocity values. The insights to model predictions provided simulate a human-level decision explanation, potentially increasing user understanding and trust.

Artificial intelligence (AI) has taken significant strides in its role as a decision-making agent, notably in self-driving vehicles. However, as AI systems become more advanced, explanations for decisions rendered by these systems may be needed to achieve a level of transparency and accountability. Towards this end, we have developed an ensemble of explainable AI (XAI) techniques, evaluated on a 1/10th scale experimental vehicle in scenarios including intersections, pedestrians, and lane changes. In advance of selecting an imminent prediction, or state identification for decision-making such as steering and speed commands, our ensemble produces a view of the path track data in an image format, with associated machine vision annotations and confidence levels. This paper presents the use of a tensor

This work is licensed under CC BY-NC-SA 4.0.

voting technique to capture this path track information, and demonstrates its application to the identification of possible imminent paths, highlighting this data in an importance map.

3.2. Challenges in Achieving Transparency

Humanizing Black Box AI relinquishes some of the levels of technical performance and robustness already achievable with current AI technologies; a compliant rationale can be contrary to optimality. Trade-offs must be made. Symbolic reasoning and logic can embody human ethics and morals more explicitly in decision-making processes, but they require expert system design. Unfortunately, these methods often suffer from reduced model performance and generalization capability. As AI is biased, the question then becomes how to ensure that AI is fair and ethically acceptable, especially in contextual decision-making scenarios. This is fundamental in the implementation of vehicle rationalized decision processes, ensuring harmony between technical innovation and moral or freedom and democracy tenets. Such trade-offs, which consider human informatic mechanisms in how rationales can be effectively visualized, prioritized and structured, can promote a better acceptance of AI systems in autonomous driving by citizens and authorities, making their use more transparent and ethical.

In advancing model transparency and interpretability in automotive AI, the utility in developing AI models which provide human-understandable rationales in critical decision-making processes is substantial. Decision-making should acknowledge and adhere to legislation, not only in terms of compliance but also to ensure that the AI decision rationale aligns with human ethics and morals. One example is the EU Regulation for vehicle type-approval, which mandates provisions for AI in braking and lane-keeping safety-related functions on M1 and N1 vehicle classes, including the provision of evidence regarding safety and transparency of the AI enabling such functions. Machine-generated rationales to justify their outputs can facilitate discussions among technical experts in various disciplines and legal professionals, which together dictate the engineering requirements, benefactors, policy makers, and corroborate broader stakeholder credibility.

4. Explainable AI Techniques

This study provides new perspectives in creating Explainable Artificial Intelligence (X-AI) for the derivation of autonomous car decision-making. The increasing use of AI models for

This work is licensed under CC BY-NC-SA 4.0.

automated car systems that perform the role of the human driver has led to concerns for access to their decision-making processes. If there is less knowledge about the operation of these AI models, people may lack trust and rely less on the models. This knowledge is more necessary, especially for situations where car accidents occur because trust in the models cannot be equivalent to that of humans. The most losses in car accidents mainly happen at intersections and three-way stop scenarios. Thus, it is crucial to predict the decision-making process of the people in those areas. For that reason, the paper employs the explanation HMI architecture to validate the state-of-the-art approaches on decision rule models, which are fitted using intersection and three-way stop scenarios data collected under the grant of the National Science Foundation.

This project will employ and compare three of the state-of-the-art methods suggested in the paper. These methods are the following: Model Explanation through Shape and Optimization (ME-SO), Model Interpretation over Joint Range of Inputs (MI-JRI), and Relation among Examples for Driverless Car's Explainability (R2DEC). ME-SO uses the information of the importance and the impact of features on the prediction score. MI-JRI shrinks down your data of instances and feature space for analysis. R2DEC utilizes the probability space generated for making a decision. The explainability techniques used in this work were obtained from an article published in peer-reviewed journals.

4.1. Interpretable Machine Learning

Interpretable machine learning is both a presently growing field and a fundamental characteristic sought in many real-world applications, especially within decision-making systems for legal and ethical reasons. In the context of autonomous driving, it is crucial to be able to trust the decisions of these systems and understand why they act as they do. Lack of trust or comprehension about the working of a complex decision-making system such as a self-driving car, which features and cues are relevant, and why a (hopefully very low) fraction of its decisions will fail, will likely result in skepticism, lack of confidence, and unpleasant surprises in novel situations. The visual nature of driving also makes it possible to use state-of-the-art neuroscience findings to better understand the cognitive processes underlying human driver decisions. The ideal would be to achieve as much transparency about the decision-making process within the vehicle as between external drivers interacting with each other, who are capable of signaling each other by making use of clearly understandable

This work is licensed under CC BY-NC-SA 4.0.

features, and eventually not following traffic and safety rules in specific critical situations in the presence of danger or external cues with clear meanings, according to state-of-the-art shared expectations about driving behavior.

Interpretable Machine Learning 4.1

In the different levels of autonomy in autonomous systems, there is a requirement of building transparent and interpretable models with quantifiable confidence and avoiding brittle failure modes. We, therefore, review relevant literature in three pivotal topics for explainable decision-making: interpretable machine learning, supervisory control and task prioritization, and uncertainty estimation.

Machine learning algorithms learn from data by extracting patterns and relationships between features and displaying these as scores, probabilities, or labels. However, when deploying AI, it is crucial to be able to interpret and explain complex models and their outcomes to ensure that model behavior aligns with social and human expectations. In the context of autonomous vehicles, designing systems that are both explainable in their decision-making and that perform adequate reasoning in critical real-world situations is of high importance. We explore current techniques to improve the transparency of autonomous system behavior beyond what supervisory regulations with real-world supervision.

4.2. Rule-based Systems

As such, how to weigh decisions or incorporate more general heuristics remains often obscure. The effectiveness of this explainability approach could break when multiple simple models are combined to provide a more powerful and accurate general on which methods, assumptions, simplifications, and approximations have been argued ad hoc to be valid. Additionally, whenever the rules are evolved during the running of an algorithm, consistency and rigor: one should inquire whether the process of creating the rules caused overfitting to specific data sets, and the applicability of the % in order to assess the model generalization capability to new, unseen scenarios. The rules should also be consistent with human expert knowledge and the model should not conflict with ethical or safety considerations.

Rule-based systems: In this type of model, decision making is characterized by a set of rules that apply to specific situations. The difficulty associated with acquiring the rules in a transparent way commonly requires some effort and domain knowledge. In cases when

This work is licensed under CC BY-NC-SA 4.0.

decision trees are used to make the decision, such rules are easily exposed and hence rule-based models give very transparent, explainable models. The simplest rule systems are characterized by if-then-else (first order) logical statements or experience-based condition-action rules and are usually human-readable. These rules can be directly understood and applied by humans, and thus can provide highly transparent decision-making. However, the laws based only on simple models may fail to capture the complexity of the environment or the idiosyncrasies of a specific scenario.

4.3. Local Explanations vs. Global Explanations

As with the loss of individual-level resolution, global explanations typically also conceal important layers of a model. An interpretable explanation or global attribute for neural networks will typically focus on the activation patterns of different neurons rather than the weight matrix which the NN optimizes. We can thus take two different perspectives. Global elements are those which contribute more to a model. From this perspective, a global explanation tries to rank the importance of different features in the model as a whole. In contrast, local explanations shed light on an individual model's decision.

Regardless of the choice of technique, explainability can also refer to different levels of coverage. We commonly differentiate between local and global explanations. Local explanations provide insight into more local and individual model predictions and clarify the reasons behind such decisions. For LIME, this in practice means creating a heuristic that comes up with explanations as to what led to the prediction made for a particular input combined with the existing data. Global explanations, instead, shed light on the inner workings of the model as a whole. The distinction between the two can be observed in the difference between opening up the observation to be explained versus looking at the model as a whole.

5. Case Studies and Applications

The efficacy of our proposed XAI techniques is first validated in closed-course conditions through rigorous comparative human driver trials. Our proposed approach is then applied to enhance the acquired autonomy algorithm to substantial safety improvements. Furthermore, a variety of driving scenarios have been developed or collected to test the transparency of the acquired autonomy algorithm. By analyzing PF_IO_IS samples of generated CNSD policies, saliency maps can be used to qualitatively assess the contribution of each PF_IS_ID to the

This work is licensed under CC BY-NC-SA 4.0.

driving actions. Experience-based explanations and programmatic explanations are also provided for other interesting scenarios to explain the behavior of the autonomous agent to human users. Our results show that the proposed method can not only facilitate human users to understand the behavior of the autonomous driving system in complex scenarios but also improve the safety of the acquired form of CNN in these complex situations in comparison with the baseline End-to-End CNN.

Case studies have been conducted to investigate the effectiveness of the proposed approach in improving the decision-making of state-of-the-art autonomous driving systems. Results show that the proposed approach substantially improved both the safety and transparency of the autonomy algorithm on a variety of safety-critical corner case scenarios. More broadly, we hope that this work will lay the foundation for the development of safer and more transparent state-of-the-art autonomous driving systems that are better equipped to deal with complex open-world environments and that human users will be more trustful and accepting of. In particular, we conduct performance studies, safety studies, and transparency studies.

5.1. Real-world Examples of Explainable AI in Autonomous Vehicles

In the context of autonomous vehicles, explainable systems have been developed. The effort to make the AI in autonomous vehicles explainable has been well captured by the supporters of AV and researchers alike. The "Explainable Robots" proposal by the European Parliament Civil Liberties Committee calls for those charged with using AI and automated vehicle technologies to implement "high explainability" to make the vehicle safer to use. Driverless transport policy, support for AI for the autonomous control of ocean-crossing sea transport has a requirement for solutions for providing real-time transparency and improve the human involvement in all phases of the ship's decision-making process. Several different machine learning techniques exist to explain the decision process. These explanation techniques are often used jointly to provide both a holistic feature summary and a holistic instance-specific view. In autonomous vehicles, there are examples of all the main explanation strands at the feature summary and feature instance level. It is important to understand the challenges and values of each explanation category.

6. Evaluation Metrics and Frameworks

This work is licensed under CC BY-NC-SA 4.0.

To establish performance bounds, we must recognize that the driving domain is not a typical reactive system. Rather, it is a system that interacts with a combination of other automatic controllers (other vehicles) and human beings. The outcome of the interaction between humans and machines depends on inputs received from the environment. Driving, therefore, is a strongly connected, in the sense of reacting to others, system. The complexity introduced by different driving laws, ethics, and personalities results in an intertwined and netted connection between driving agents. Identifying optimization points in such a complex system and formulating appropriate individual and collective objective functions and constraints is a very complex problem. In traffic, autonomous vehicles interact not only with each other but also with "intelligent" human drivers. Therefore, an autonomous vehicle not only has to optimize its individual desirable driving objectives but also needs to be aware of and adapt to the combined decision-making process of other autonomous vehicles. Designing such interactions to nudge traffic towards global optimization is a complex problem.

This chapter highlights the pressing need for broader and more detailed evaluations of driving systems, with a consideration of their interaction with humans. Safety evaluation metrics are, in general, based on statistical comparisons of recorded observations to normative behaviors. Few metrics incorporate driver models or other interaction or usability results. We argue that these represent incomplete evaluation frameworks and point in the direction of more sociotechnically-oriented ones. For intelligent autonomous systems that interact or interface with human beings, or are responsible for safety-critical tasks, it is essential to develop, in addition to the existing algorithmic verifications, another level of verification that addresses the broader decision- and information-centric aspects of intelligent systems. What kind of QA does society expect from AI systems, beyond functionality and verifiability? What types of demonstration does the public embrace? Should legal standards be established?

6.1. Accuracy vs. Interpretability Trade-off

AI models must give explanations why they think a particular output is correct. These explanations are an integral part of most AI deployment. For example, decisions about loans, judicial and medical diagnoses. Therefore, we must build AI systems that are able to give explanations for their outputs on certain tasks.

In regions of highly accurate models, when the model's outputs are intuitive, and in cases, when the output can be verified, the user can rely on them. The main weakness of AI models

This work is licensed under CC BY-NC-SA 4.0.

is that they are an inaccurate statistical reflection that may be incompatible with the user's intuitions or available data. Algorithms work with hidden factors that remain unobserved makes data unavailable for explanatory analyses. AI offers accurate predictions but not complete understanding. Users must weigh whether they can trust a model's accuracy and whether they will get into trouble either using the model or not using it. Answers to these questions are a decision-making process vital.

For example, in computer vision, a model that monitors security footage in real-time will be predicated on fidelity or accuracy since any reduction in performance could allow a breach that could compromise the entire system and perhaps the organization.

A model that is the most accurate or has higher classification accuracy is the most preferred by the end-user. In some application scenarios, model accuracy is paramount, and interpretability is a secondary consideration. The user of the model is not as concerned about the model's internal structure but is interested only in its decision and cannot be bothered with details of how the model arrived at that recommendation.

Discuss the traditional accuracy vs. interpretability trade-off, followed by a discussion of Explainable AI or XAI techniques that have been proposed to address this trade-off.

6.2. Fairness and Bias in Explainable AI

The simplest approach is to consider aggregating or embedding biases or fairness concerns in the model output, for example an additional output or counterfactual result or else downstream predictive inference completely focused on fairness or bias specifically. We could imagine a variant of a different kind of reward or loss function that consists of a weighted combination of the original utility or decision making function and a separate function that measures a breakdown of the fairness criterion or lacks of group-level parity. A low value of this fairness constraint would be a desired trait of the AI explainability aspect that provides support for decision making governance, potentially based on individual and aggregated user feedback excluding sensitive personal information or relying on outcomes from group discussions.

As the concept of fairness gains increasing importance in sophisticated machine learning models and larger-scale AI systems, it is natural to include this design criterion in the model explainability process. Many aspects of model fairness and bias can be addressed in existing

This work is licensed under CC BY-NC-SA 4.0.

machine learning techniques, and when the results or reasoning provided by a model are to be communicated to lay or non-expert stakeholders, the definitions of fairness that are used to request data that is representative of these concerns. General consensus opinion among these stakeholders or a regulatory standard or industry initiative that sets guidelines on these practices can, therefore, help to guide the research focus in XAI with regard to fairness or bias.

7. Regulatory and Ethical Considerations

Finally, moral dilemmas are one of the most complicated questions in designing decision-makers responsible for the operation of vehicles in the presence of potentially dangerous situations. Several principles were suggested in the current literature of Responsibility Sensitive Safety (RSS) in respect to three different fronts that impact ethical dilemmas in avoiding situations on the part of autonomous vehicles.

It is important for self-driving cars to demonstrate safe operation to gain public trust and regulatory acceptance. When an accident occurs, the mechanism responsible for the driving must have a high level of responsibility to avoid further damage. Car manufacturers will have to design systems considering not only technology but also legislation and ethics preferences at a local level. The design of the system and the decision-making model will be strongly impacted by localization.

An important societal concern is the potential impact on the labor market, as millions of people in many countries work using motor vehicles. The career of these people may be affected by the increase in the number of vehicle automation levels. The introduction of self-driving cars may bring benefits for people who cannot drive, such as disabled and elderly individuals. The vehicle plays the driver's role, allowing these people greater autonomy.

Increasing automation leads to questions of liability. In a highly autonomous vehicle, the driver is a passenger and may not require a driving license. Therefore, new liability allocation and insurance rules will become necessary. The driver is also not required to be competent in controlling the vehicle to deal with abnormal situations due to the operational design domain (ODD) of the vehicle. When problems with the system itself occur, the driver is also not able to fix the issue. The vehicle requires a robust solution by design to guarantee reliability and safety.

7.1. Current Regulations and Guidelines

This work is licensed under CC BY-NC-SA 4.0.

Building these conversations with a broader set of stakeholders and evaluating the multiple trade-offs implied by the various motivations and constraints in terms of designing explainability in AI can lead to the production of more robust and acceptable AV. This promise of meaningful and explainable AI development in the field of autonomous vehicles is the main topic of discussion and inquiry in our article and is based on insights regarding value-oriented conversations with French AV companies.

These regulations and guidelines are applied differently by regulators across the world. Here, we focus on general guidelines and regulations issued by regulators and individual companies that manufacture L3 and L4 autonomous vehicles. Tackling the challenge of developing an acceptable set of explainable AI devices that comply with legal regulations and safety standards, and that meet stakeholder expectations, processes represent a watershed moment for the promising field of explainable AI and for the community of autonomous-vehicle (AV) developers.

7.2. Ethical Implications of Explainable AI in Autonomous Vehicles

Both XAI-ACH and XAI-PH have to actively adhere to what we will name the multi-purpose, ACH-advocating function of maximal, i.e. most comprehensive transparency that S. Wachter, B. Mittelstadt, and C. Russell illuminated: Explainable AI can help us understand how much decision-making authority we otherwise unknowingly concede to these systems. Failure to provide and display these promised capacities for providing explanations indicates a lack of trust in the manufacturers of these systems and of their alleged lets to constitutes and planners. Understanding their consequences implies an understanding of the decision-makers involved in the analysis and of the societal governance that is added in by all those who have a status to address the company or any other benefactor who formerly enabled the design and deployment of these systems. To understand this style of decision has to be provided with varying levels and the sort of information most suitable for that person's reasoning processes.

The deployment of XAI in AVs is associated with potential ethical implications, like any other technology. By clarifying why a consequence has emerged, clarification could enhance human experience about a fatal AV accident and lessen the anguish of people profoundly disturbed by the incident. Finding appropriate deployment areas for XAI for AVs, based on several dimensions of transparency that target distinct human agents, their beliefs, and reasoning styles, can save following assessments and helpful behaviors—considering that AV-related

This work is licensed under CC BY-NC-SA 4.0.

features do not wholly depend on which ultra-low latency, full autonomy-making AVs completely hidden computational classifiers ultimately enforce. The same vision can also be used as a guide for the de-selection of XAI methods that circumvent facts that are considered relevant for decision-maker appraisal and perception of their algorithmic AVs. We hope to initiate discussions about these dimensions of ways their methods dodge them in real-world AV deployment and create axioms that point out preferable aversion strategies for highly autonomous vehicles.

8. Future Directions and Emerging Trends

1.2 Explanations Focused on the Regret There is recent work in the areas of human-computer interaction and recommender systems that emphasizes the need for systems to explain with more than just heuristics. The system should provide more generic reasons for arriving at the decisions it does. It is now seen as important to 'improve prediction quality, where it is needed, for cases of high user uncertainty', and 'predictive models that can provide better explanations for potentially suboptimal predictions by balancing quality of explanations and prediction accuracy allow for controlled decision making. The notion of quality of explanations must also be extended from simplistic measures like how accurate the explanation of a prediction or classification made by the system is, to more intuitive and domain-relevant concepts like regrets. In this work, the explanation while generated posthoc of a machine learning or inference decision shows the criteria that caused the particular suggestion for a data instance to be an undesirable suggestion, and how that instance could have been modified to get a prediction with smaller overall regret. Model explanatory functions must minimize the discrepancy between an ideal model and the learned model while consistently capturing desired qualitative model features like relevance, clarity, and fairness.

The need for achieving transparency and interpretability in decision-making has been evoked by researchers and does not find a limit in the universe of techniques that can be employed. We list a couple of illustrative directions that we think will be receiving more attention in the near future, especially in applications like autonomous vehicles, which demand very high reliance by the users or where the latency created by delays due to algorithmic explanations cannot be entertained in the decision-making chain.

8.1. Advancements in Explainable AI Research

This work is licensed under CC BY-NC-SA 4.0.

Nonetheless, there exist many promising techniques, algorithms and challenging exploration ideas towards evaluating and improving our knowledge on AI systems, including autonomous decision-making and XAI is considered of high importance in the current AI roadmap of both academia and industry. There is a growing demand for implementation of policy and regulations ensuring the development of the responsible AI system, where the AI decision-making algorithm provides transparency and is subject to a thorough evaluation. Today, for example, many financial institutions have regulatory requirements to implement explanatory models in AI-based credit-decision support tools and in the European Union's General Data Protection Regulation (GDPR), the right to explanation is becoming a fundamental right of the autonomous decision-making algorithms, which directly affects the machine learning community.

Even though XAI research, particularly focused on increasing transparency in decision-making, is not new, there is a growing number and variety of XAI techniques. However, to date, no single method used for high-stake decision-making is a reliable solution, and many of the developed techniques do not address the particular challenges in developing an algorithmic transparency for an autonomous driving system. For many techniques, their explanation algorithms judged over more accurate models, such as deep learning models, are computationally expensive, hard to train or require approximations, where correctness and incompleteness to provide explanations become another reliability challenge. In addition, using explanations effectively is also a problem since many users tend to not understand the given explanation and do not trust the decisions being made.

The subject of Explainable AI (XAI) is gaining significant attention, as many practical AI systems are black-boxes in terms of how they make decisions. There is a growing concern that the black-box nature of AI systems cannot be trusted, and people need some kind of explanation that is comprehensible and transparent. This is especially important when it comes to AI applications used in critical sectors such as healthcare, judicial systems, defense systems, and autonomous vehicles. The researchers in XAI are highlighting that transparency is not only getting an explanation, but it is also providing confidence, understanding, and capability to use the AI system effectively.

9. Conclusion

This work is licensed under CC BY-NC-SA 4.0.

Given the urgency in employing AI in autonomous vehicles, explainable AI techniques have received less attention. Alghamdi et al. propose that combining explainable AI techniques with usable design principles such as interface abstraction and feedback can help the user comprehend and trust a self-driving car. Motorists are likely to comply more with the output of a decision model that generates confident and explainable outputs, and thus the explainable AI models can result in reduced human feedback requirements. In addition to increasing user trust and satisfaction, better user compliance in self-driving cars increases the safety of the AI system.

The goal of this chapter was to address ethical concerns in AI systems with respect to ensuring transparency in decision making. The use of more sophisticated machine learning models in AI agents often leads to the systems producing high discrimination between user requested output and the model predicted outcome. Increasing transparency has been shown to significantly reduce this discrepancy, as well as result in generally better performance of the AI agent, lower human feedback requirements, and the system being perceived as more trustworthy. We focused the discussion on autonomous vehicles with a view to increasing transparency of the decision-making process when we approach them for user assistance. We proposed explainable AI techniques that can make the decision-making process of such AVs more transparent to the user.

9.1. Summary of Key Findings

In terms of the tools that could be developed to address any need for these explanatory models to be useful within development activities, we have discussed how interactive examples could be used to support explanations and how these could be integrated into other kinds of test activities. We hope the chapter provides a useful guide for practitioners in the field with respect to understanding which attributes of black-box systems can potentially be addressed using which kinds of techniques. We also hope the examples provide a sufficient baseline for practitioners in the field to decide these are largely appropriate kinds of approaches to suit the specific needs of autonomous vehicle development in promising directions. Overall, the potential of these techniques when designing high-value future systems, combined with the early adoption possibilities provided by the close association of development validation with assurance regimes, suggests promising development paths for the application of

This work is licensed under CC BY-NC-SA 4.0.

development, test, and validation activities as ever-larger development projects start to adopt them.

This chapter has provided an overview of a number of techniques from explainable AI research that have particular promise for use in the domain of autonomous vehicles. We have aimed to provide a number of examples of techniques that span a variety of the different components and complex tasks performed by autonomous vehicle designs with a focus on the potential uses in development processes. We provided within-subsystem examples for perception, behavior prediction, decision making, and motion planning and discussed examples of cross-subsystem approaches targeted at developing models which can explain actions across multiple systems. In this discussion, we aimed to highlight practical considerations, particularly those relating to the models' potential to support development, validation, testing, and risk assessment activities for autonomous vehicle deployment.

9.2. Implications for the Future of Autonomous Vehicles

This chapter discusses the future of the autonomous vehicle guided by the development process of ethical issues explained in the chapter. In particular, transparency of decision-making processes is highlighted. Insufficient level of transparency of decision-making processes could drive users from adopting AVs. In general, many autonomous systems will face an ethical or moral barrier without an insight into AI decision mechanisms. Although technology is developed rapidly, the introduction of AI systems into society, including autonomous vehicles, depends on the public's understanding and acceptance of them. This chapter states we need to enhance the transparency, controllability, and ethical considerations for the decision-making mechanisms of the autonomous vehicle for its future acceptance. Some possible AI methods for meeting these requirements are considered. Finally, this chapter also introduces implications from the technical development viewpoint.

Although traditional AI models are typically easily interpretable and their decision-making processes can be traced, many new applications come with a complex black-box model such as deep learning algorithms. As shown in the previous section, autonomous vehicles (AVs) ethics heavily depend on the AV's decision mechanism and decision-making process. The heavy use of complex black-box models jeopardizes the AV's adoption. Meanwhile, it is important to shift our viewpoint away from the physical accident towards the nature of the decision-making process of the autonomous vehicle. Even when the society accepts a risk of

This work is licensed under CC BY-NC-SA 4.0.

traffic accidents associated with AV adoption, no one will take an AV which provides decisions that could not be understood by others. Ethical implications and social acceptance are closely connected and influence each other. Roads should be safer when more people use AVs. This chapter discusses some ethical implications concerning the future use of the autonomous vehicle.

10. References

1. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1955-1969, May 2018.
2. R. Caruana et al., "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia, 2015, pp. 1721-1730.
3. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135-1144.
4. A. L. Mille, G. Lena, and F. Yvon, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019.
5. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 3338-3347.
6. S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4765-4774.
7. T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019.

This work is licensed under CC BY-NC-SA 4.0.

8. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 3319-3328.

9. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, 2013.

10. Tatineni, Sumanth. "Cloud-Based Business Continuity and Disaster Recovery Strategies." *International Research Journal of Modernization in Engineering, Technology, and Science*5.11 (2023): 1389-1397.

11. Vemori, Vamsi. "From Tactile Buttons to Digital Orchestration: A Paradigm Shift in Vehicle Control with Smartphone Integration and Smart UI–Unveiling Cybersecurity Vulnerabilities and Fortifying Autonomous Vehicles with Adaptive Learning Intrusion Detection Systems." *African Journal of Artificial Intelligence and Sustainable Development*3.1 (2023): 54-91.

12. Shaik, Mahammad, Leeladhar Gudala, and Ashok Kumar Reddy Sadhu. "Leveraging Artificial Intelligence for Enhanced Identity and Access Management within Zero Trust Security Architectures: A Focus on User Behavior Analytics and Adaptive Authentication." *Australian Journal of Machine Learning Research & Applications* 3.2 (2023): 1-31.

13. Tatineni, Sumanth. "Security and Compliance in Parallel Computing Cloud Services." *International Journal of Science and Research (IJSR)* 12.10 (2023): 972-1977.

14. J. C. Carvalho et al., "Explainable Artificial Intelligence for Predictive Maintenance," *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, Rhodes, Greece, 2020, pp. 1-6.

15. J. J. A. Pereira et al., "Explainable Artificial Intelligence for the Prediction of Depressive Symptoms in Older Adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 724-732, 2020.

16. A. Stumpf, "Interpretability, or How to Explain AI Models," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 24, no. 3, pp. 32-35, 2018.

This work is licensed under CC BY-NC-SA 4.0.

17. S. D. Jain, A. Agarwal, and S. K. Dhurandher, "Explainable AI: A Review," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-6.

18. T. N. K. Venkata and M. V. Malathy, "Interpretable Machine Learning Models for Healthcare," *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2020, pp. 1-6.

19. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

20. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1955-1969, May 2018.

21. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 3319-3328.

This work is licensed under CC BY-NC-SA 4.0.