Explainable Artificial Intelligence in Autonomous Vehicles: Methodologies, Challenges, And Prospective Directions

RAPHAEL UGBOKO[1], OLUWAFEMI OLORUNTOBA[2]
[1]*Human-Centered Computing, Clemson University, USA*
[2]*Department of Information Technology, Lamar University, USA*

Abstract- ***The increasing complexity of autonomous vehicle (AV) decision-making systems driven by deep learning and black-box models has intensified the need for explainable artificial intelligence (XAI). This paper explores the integration of XAI within AV systems, focusing on methodologies that enhance interpretability without compromising real-time performance and safety. We provide a structured taxonomy of XAI approaches, comparing post-hoc techniques such as LIME and SHAP with inherently interpretable models like decision trees and linear classifiers. The paper also investigates causal reasoning, human-machine trust, ethical concerns, and regulatory implications. Through analysis of current challenges and emerging solutions including inherently interpretable neural networks and standardized XAI benchmarks we offer a roadmap for future research. Our findings underscore the critical role of XAI in fostering trust, accountability, and safe deployment of autonomous systems.***

Indexed Terms- ***Explainable Artificial Intelligence (XAI), Autonomous Vehicles (AVs), Model Interpretability, Human-AI Trust, Safety-Critical AI Systems***

I. INTRODUCTION

1.1 Contextualizing Explainable AI in Autonomous Vehicle Systems

Autonomous vehicles (AVs) represent a significant advancement in transportation, holding potential for enhanced safety and efficiency (Zöllner & Schamm, 2015). These systems rely on sophisticated artificial intelligence (AI) algorithms for perception, decision-making, and control (Minaie et al., 2020)(Wylde, 2012.). As AVs become more prevalent, the opacity of their AI models, often referred to as 'black boxes', presents considerable challenges (Padl, et al., 2020). This lack of transparency impedes understanding why an AV makes a particular decision, especially in critical situations like accident avoidance (Betz et al., 2019) (Vida & Váradi, 2018). Explainable Artificial Intelligence (XAI) addresses this by rendering AI system outputs comprehensible to human users (Larasati & Deliddo, 2020).

The integration of XAI into AV systems is therefore crucial for fostering trust, ensuring accountability, and facilitating regulatory compliance (Hakimi, 2018) (Leslie, 2019). Without explanation, stakeholders cannot ascertain if AV's decisions align with safety protocols, ethical considerations, or legal frameworks (Krontiris et al., 2020) (Michael et al., 2020). This paper examines the methodologies, challenges, and future trajectories for XAI in the context of AV development and deployment.

A high-profile incident involving Uber's self-driving car in 2018, where a pedestrian was fatally struck, underscored the urgent need for transparency in AV decision-making. The inability to trace or interpret the vehicle's internal logic during the critical seconds preceding the crash illustrates the potential consequences of opaque AI behavior. Such cases highlight the imperative for robust, real-time explainability mechanisms in AV systems.

1.2 Research Objectives, Scope, and Significance

The objective of this research is to systematically review the current state of XAI in autonomous vehicles. Specifically, this paper:

- Identifies prevailing methodological paradigms for achieving explainability in AV systems.
- Examine the trade-offs between model interpretability and performance.
- Explores the impact of XAI on human-machine interaction and user trust.
- Analyzes safety, ethical, and regulatory considerations pertinent to XAI implementation in AVs.

The scope encompasses AI models used for perception, planning, and control within AV architecture. This analysis excludes general AI explainability techniques not directly applied or adaptable to the AV domain. The significance of this work is due to its comprehensive overview of XAI in a high-stakes application area. It provides a foundation for future research and development, addressing critical issues of transparency and accountability necessary for widespread AV adoption (Schäbe, 2019).

1.3 Background: AI Architecture in Autonomous Vehicles

AV perception–prediction–planning pipeline

The AV perception–prediction–planning pipeline typically integrates multimodal sensor data to construct environmental models, enabling real-time object detection, tracking, and trajectory forecasting for downstream decision-making modules. Recent advancements employ deep learning architectures, such as convolutional and recurrent neural networks, to enhance the accuracy and robustness of perception and prediction tasks within this pipeline (Padl, et al., 2020).

These deep learning-based models, while effective, often operate as opaque systems, making it difficult to trace the reasoning behind their outputs and decisions in real-world driving scenarios (Massoud & Laganiere, 2024). As AV pipelines integrate increasingly complex neural architectures, the necessity for explainability becomes more pronounced to ensure both operational transparency and system reliability (Pavel et al., 2022).

Recent research has introduced explainability techniques tailored to deep learning models within AV pipelines, such as post-hoc saliency mapping and feature attribution methods, to elucidate model behavior in complex scenarios (Patel et al., 2021). Additionally, the integration of uncertainty quantification alongside explainability offers a pathway to enhance safety validation and regulatory acceptance of AV decision-making processes (Patel et al., 2021).

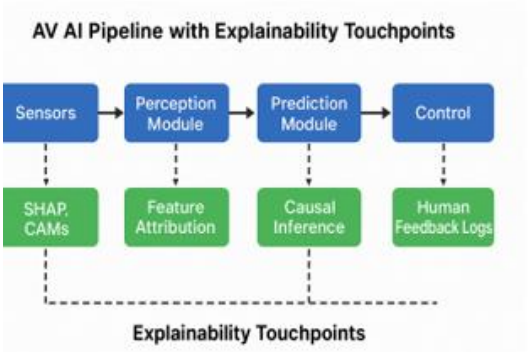

Figure 1: AV AI Pipeline with Explainability Touchpoints

Figure 1 above visualizes the standard AV decision-making pipeline and highlights where explainability techniques can be integrated. Each module can be made interpretable using different XAI methods based on its data type and output structure.

Where explainability applies across the pipeline

Explainability is relevant not only for perception modules but also for prediction and planning stages, where model transparency can clarify the rationale behind trajectory forecasts and maneuver selections. Interpretable outputs in these downstream modules are critical for diagnosing errors, validating safety constraints, and supporting regulatory review of AV decision-making processes (Nazat et al., 2024) (Teeti et al., 2022).

Moreover, explainability techniques are increasingly being adapted for complex prediction and planning modules, such as attention-based mechanisms and counterfactual reasoning, to clarify model reasoning in uncertain or multi-agent environments (Limeros et al., 2022). Integrating these approaches enhances post-hoc interpretability and supports the identification of failure modes in safety-critical AV scenarios (Omeiza et al., 2022).

II. METHODOLOGY

2.1 Data Collection Strategies and Dataset Description

This research employs a systematic literature review approach to gather relevant studies on XAI in autonomous vehicles. Publications were identified through searches across major academic databases, including IEEE Xplore, ACM Digital Library, Scopus, and Web of Science. Search queries combined terms such as "explainable AI," "XAI," "interpretability," "transparency," "autonomous vehicles," "self-driving cars," "automated driving," and "driverless cars." The selection criteria prioritized peer-reviewed journal articles, conference papers, and reputable preprints published within the last five years to ensure currency. Inclusion extended to works detailing specific XAI methods applied to AV tasks, discussions on the challenges of explainability, and analyses of regulatory or ethical implications.

The dataset for this review comprises a diverse collection of theoretical frameworks, empirical studies, and conceptual analyses. For instance, several papers detail the application of machine learning (ML) models in AVs, ranging from perception tasks like object detection to complex decision-making processes (Gao et al., 2019) (Schwarting et al., 2019). Other sources discuss the need for robust training datasets and the potential biases within them, which are directly relevant to XAI's goal of understanding model behavior (Ahmad Fawad, 2023). The literature also includes qualitative analyses of public perception and trust in AVs, providing crucial context for human-centric XAI requirements (Pettigrew et al., 2019) (Zhang et al., 2020).

2.2 Analytical Frameworks and Evaluation Metrics

The collected literature was analyzed using a thematic synthesis approach. This involved iteratively identifying recurring themes, concepts, and arguments related to XAI in AVs. Three primary analytical frameworks guided the review:

Technical Explainability: Focuses on methods that reveal the internal workings of AI models, such as feature importance, rule extraction, or counterfactual explanations.

Human-Centric Explainability: Examines how explanations are presented to users, their effectiveness in building trust, and their impact on user understanding and acceptance (Haspiel et al., 2018).

Societal and Regulatory Explainability: Considers the broader implications of XAI for legal liability, ethical decision-making, and policy development (Žolnerčíková, 2019)(Anderson et al., 2016).

Evaluation metrics for XAI vary depending on the specific objective. For technical explainability, metrics include fidelity (how well the explanation reflects the model's behavior), interpretability (ease of understanding for a human expert), and stability (consistency of explanations for similar inputs). For human-centric aspects, metrics often involve user studies assessing trust, satisfaction, and task performance with explanations. Regulatory compliance metrics are less standardized but generally involve adherence to principles like transparency, fairness, and accountability (Leslie, 2019). A key challenge remains the lack of universally accepted benchmarks for XAI in the AV context, often necessitating qualitative assessments alongside quantitative measures (Ahmad Fawad, 2023).

To assess the effectiveness of XAI techniques in AV contexts, we draw upon a combination of technical and human-centered metrics. Table 1 summarizes the most relevant evaluation dimensions across fidelity, latency, trust, and interpretability.

Table 1: Key Evaluation Metrics for XAI

Metric	Description	Type	Relevance to AVs
Fidelity	Agreement between explanation and model prediction	Quantitative	High (accuracy of rationale)
Comprehensibility	How easily a human can interpret the explanation	Qualitative	Critical for driver override
Latency	Time cost of generating explanations	Quantitative	Essential in real-time systems
Simulatability	Whether a human can mentally simulate the model's reasoning	Qualitative	Affects trust and acceptance
Completeness	Extent to which explanation captures total model behavior	Quantitative	Needed for auditing
Human Trust Score	User-rated metric on perceived reliability after explanation	Empirical	Core to deployment usability

Table 1 outlines the core evaluation metrics used to assess the quality and utility of XAI techniques in autonomous vehicle (AV) systems. It distinguishes between quantitative, qualitative and empirical dimensions—covering both technical performance (e.g., fidelity, latency, completeness) and human-centered factors (e.g., trust score, comprehensibility). These metrics are essential for understanding the trade-offs between transparency, usability, and computational efficiency in real-world AV applications.

2.3 Benchmark Datasets and Evaluation for XAI in AVs

Notably, widely used benchmark datasets such as KITTI and nuScenes have been leveraged to evaluate XAI techniques within AV pipelines, enabling standardized assessment of both perception and decision-making modules (Sajjad et al., 2022)(Wickramarachchi et al., 2024). Recent studies also highlight the importance of developing domain-specific XAI evaluation protocols that account for the unique operational requirements and safety constraints of autonomous driving scenarios (Gyevnar et al., 2025).

Emerging explainability methods now extend to multi-modal sensor fusion tasks, where techniques such as attention visualization elucidate how AVs integrate lidar, radar, and camera data for robust perception and decision-making (Dong et al., 2023). Additionally, recent datasets like DeepAccident enable direct, explainable evaluation of accident prediction models, supporting transparent assessment of AV safety in complex scenarios (Wang et al., 2023).

Emerging research highlights the integration of interpretable attention mechanisms and uncertainty quantification within AV pipelines, which enhances both the transparency and reliability of trajectory prediction and decision-making modules (Wu et al., 2023)(Li et al., 2024). These advancements facilitate more robust validation processes and support clearer regulatory assessments of AV system behavior in complex, real-world environments.

Effective evaluation of explainability methods requires diverse, well-annotated datasets. Table 2 compares prominent AV datasets in terms of their sensor coverage, label richness, and suitability for benchmarking XAI techniques.

Table 2: AV Datasets and Explainability Potential

Dataset	Sensors	# Scenes	Scenario Types	Label Richness	Use in XAI Studies
KITTI	Camera, LiDAR	~14,000	Urban, Highway	Medium	Yes (SHAP, Grad-CAM)
nuScenes	Multi-modal	~1,000	Weather, Night	High	Yes (LIME, Anchors)
Waymo Open	Multi-modal	~20,000	Real-world traffic	Very High	Yes (Visual XAI)
Argoverse	LiDAR, Maps	~3,000	Complex turns, merges	Medium	Limited

Table 2 compares major open-source AV datasets in terms of their sensor configurations, diversity of scenarios, label richness, and suitability for evaluating XAI methods. It supports dataset selection by researchers aiming to benchmark explainability in perception and decision-making models under various traffic and environmental conditions.

III. THEMATIC LITERATURE REVIEW OF EXPLAINABLE AI IN AUTONOMOUS VEHICLES

3.1 Methodological Paradigms: Approaches to Explainability in AV Systems

Various methodological paradigms address explainability in autonomous vehicle systems. One prominent category involves post-hoc explainability techniques, which generate explanations after an opaque AI model has made a prediction. Examples include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which approximate the behavior of complex models locally or globally to provide feature importance scores. These methods are particularly useful for deep learning models that inherently lack transparency (Grossberg, 2020). For AVs, this could mean identifying which sensor inputs (e.g., camera data, lidar readings) or environmental factors primarily influenced a specific driving maneuver (Betz et al., 2019).

Various XAI techniques offer different strengths and trade-offs. Table 3 provides a comparative overview of key post-hoc and inherently interpretable methods used in AV systems, highlighting their fidelity, computation cost, and application domains.

Table 3: Comparison of XAI Techniques for Autonomous Vehicles

Technique	Type	Fidelity	Computation Overhead	AV Use Cases	Pros	Cons
LIME	Post-hoc	Medium	Medium	Decision justification for planning	Local interpretability	Fragile to perturbations
SHAP	Post-hoc	High	High	Visual perception, trajectory prediction	Global + local fidelity	High latency
Grad-CAM	Post-hoc	Medium	Low	Image classification in perception	Visual heatmaps	Limited to CNNs
Anchors	Post-hoc	Medium	Medium	Rule-based explanations	If-then rules, intuitive	Sparse explanations
DT	Inherent	High	Low	Route planning, obstacle prioritization	Fully interpretable	Poor scalability

Table 3 table offers a side-by-side comparison of XAI techniques commonly used in AV systems, distinguishing between post-hoc and inherently interpretable models. It evaluates each technique based on fidelity, computational overhead, typical

AV use cases, and practical advantages and disadvantages

Another paradigm is inherently interpretable models, where the model's structure allows for direct understanding of its decision-making process. Decision trees and linear models fall into this category. While offering high transparency, these models often struggle to achieve the same level of performance as complex deep neural networks, especially in the high-dimensional, real-time environments of autonomous driving (Grossberg, 2020). Hybrid approaches, combining simpler, interpretable models with more complex, high-performing components, also represent a developing area. For instance, an AV's path planning might use an interpretable rule-based system for critical decisions, while a neural network handles perception tasks.

Furthermore, causal inference and symbolic AI methods are gaining traction for their ability to provide human-understandable reasoning. Causal models aim to identify cause-and-effect relationships, offering explanations that go beyond mere correlations. Symbolic AI, leveraging knowledge representation and logical reasoning, can generate explanations in natural language or as logical rules, mirroring human cognitive processes. For AVs, this could mean articulating a decision like "I applied brakes because a pedestrian entered the crosswalk" with explicit cause-and-effect links. The challenge with these approaches often lies in their scalability and ability to handle the uncertainties and complexities of real-world driving scenarios (Schwarting et al., 2019).

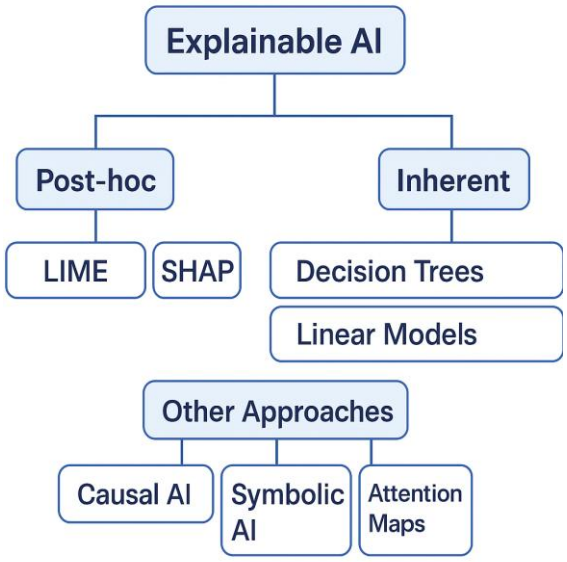

Figure 2: Taxonomy of XAI Methods in Avs

Figure 2 presents a taxonomy of XAI techniques categorized into post-hoc and inherently interpretable methods, along with emerging hybrid approaches relevant to AVs.

3.2 Interpretability Versus Performance: Trade-offs in Model Design

A fundamental tension exists between the interpretability of an AI model and its performance, particularly accuracy, in the context of autonomous vehicles. Models that achieve state-of-the-art performance in perception and control tasks, such as deep neural networks, are typically highly complex and non-linear, rendering their internal decision processes opaque (Padl, et al., 2020) . Conversely, models that are inherently interpretable often cannot capture the intricate patterns and nuances required for robust AV operation in dynamic and unpredictable environments (Ahmad Fawad, 2023).

This trade-off necessitates careful consideration in AV design. In safety-critical components, a higher degree of interpretability might be prioritized, even if it means a slight reduction in peak performance, to ensure verifiability and accountability (Schäbe, 2019)(Rokseth et al., 2019). For example, the core decision logic for emergency braking might be designed to be fully transparent. However, for perception tasks like object recognition, where deep learning excels, some level of opacity might be

accepted, with explainability techniques applied post-hoc to justify specific classifications. The challenge is to find the optimal balance where explanations are sufficient for human oversight and trust, without unduly compromising the vehicle's operational capabilities (Hakimi, 2018). Research explores methods to mitigate this trade-off, such as developing "glass-box" models that offer both high performance and inherent transparency, or by integrating XAI techniques directly into the training process.

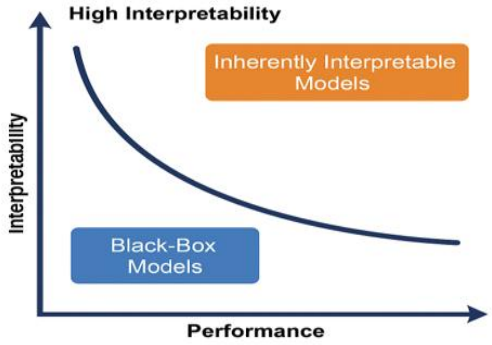

Figure 3: Explainability-Performance Trade-off in Model Design

Figure 3 shows Models with higher interpretability often sacrifice some accuracy, and vice versa.

3.3 Equations

(a) Fidelity of Explanation

Measures how well the explanation model approximates the original:

Fidelity(E,M)=1-1/n ∑_(i=0)^n▒|M(x_i)-E(x_i)|

- Description: E is the explanation model, M is the original model, x_i are input samples. This reflects the faithfulness of the surrogate explanation.

Here, M represents the original model, E is the explanation model, andx_i is the input instance. A lower value indicates a closer match between the explanation and the model's actual output.

(b) Trust Metric with Multi-Factor Composition

T=α·U+β·C+γ·L

- Description: A linear model to represent Trust T in AVs. U = User Satisfaction, C = Comprehensibility, L = Latency. Weights α,β,γ are dependent on context.

T denotes trust, calculated as a weighted sum of user satisfaction (U), comprehensibility (C), and latency (L), with tunable weights α,β,γ.

(c) Optimization Formulation: Accuracy vs Interpretability

(_M^min)L(M)+∫_0λ·Ω(E)

Description: Optimization objective balancing prediction loss L and complexity of explanation Ω(E), where λ tunes the trade-off.

This loss function balances model accuracy L(M) and explanation complexity Ω(E), controlled by the trade-off parameter λ.

3.4 Real-World Case Studies on XAI in AV Development (e.g., Waymo, Tesla)

For example, Tesla's Autopilot system has incorporated visualizations and driver feedback mechanisms aimed at increasing user understanding of AI-driven maneuvers, while Waymo has explored scenario-based explanations for complex navigation decisions (Kumari & Bhat, 2021)(- et al., 2024). These real-world deployments highlight both the technical feasibility and the ongoing challenges of achieving actionable explainability in safety-critical, real-time environments. Recent deployments have also begun leveraging federated learning and distributed ledger technologies to enhance traceability and auditability of AI-driven decisions in AVs, thereby supporting regulatory transparency and data provenance (Padl, et al., 2020). Furthermore, the integration of trusted execution environments enables secure, privacy-preserving explainability mechanisms that can scale to complex, data-intensive AV scenarios (Padl, et al., 2020).

Recent advancements also include the application of distributed ledger technology (DLT) and trusted execution environments (TEEs) to enhance the traceability and auditability of AI-driven decisions in AVs, supporting secure and privacy-preserving explainability mechanisms (Padl, et al., 2020). These

approaches facilitate the creation of immutable records for model metadata and data flows, thereby strengthening accountability and regulatory compliance in autonomous vehicle deployments (Padl, et al., 2020).

3.5 Human-Machine Interaction and Trust: User-Centric Perspectives

The successful adoption of autonomous vehicles is intrinsically linked to public trust, which XAI significantly influences (Hakimi, 2018) (Pettigrew et al., 2019). Users, whether drivers, passengers, or pedestrians, need to understand and predict an AV's behavior, especially in unexpected or ambiguous situations (Cunneen et al., 2019). XAI facilitates this by providing intelligible explanations for decisions, actions, or failures, thereby building a mental model for the user of how the AV operates. Studies indicate that explanations provided before an autonomous action enhance trust more effectively than explanations given after the fact (Haspiel et al., 2018).

User-centric XAI focuses on the presentation format and content of explanations to cater to diverse user needs and cognitive abilities. Explanations can take various forms, including visual cues (e.g., highlighting sensed objects), textual descriptions (e.g., "Yielding to pedestrian"), or auditory warnings. The level of detail and complexity must be adaptable; a general passenger might require a high-level summary, while a safety engineer would need a detailed technical breakdown. Miscalibrated trust, either over-trust or under-trust, poses risks. Over-trust can lead to complacency and reduced vigilance, while under-trust can result in disuse or manual overrides that diminish the benefits of automation. Effective XAI aims to foster appropriate trust, aligning user expectations with system capabilities (Zhang et al., 2020). Communication of intent and future actions is a critical component of this interaction (VINKHUYZEN & CEFKIN, 2016).

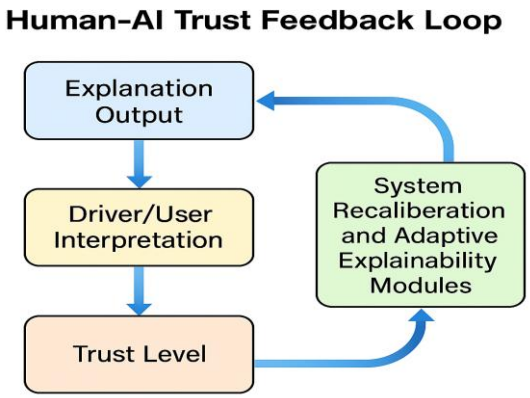

Figure 4: Human-AI Trust Feedback Loop

Figure 4 demonstrates the human-in-the-loop paradigm where explanations build or erode user trust, which in turn informs system design and XAI deployment policies.

3.6 Safety, Ethics, and Regulatory Considerations in XAI Implementation

The implementation of XAI in autonomous vehicles is deeply intertwined with safety, ethical, and regulatory considerations. From a safety perspective, XAI can serve as a crucial tool for validation and verification, allowing engineers to debug and refine AV algorithms by understanding the root causes of errors or unexpected behaviors (Betz et al., 2019). This is particularly relevant given the potentially catastrophic consequences of AV failures (Filiz, 2020). Explanations can also assist human operators or remote supervisors in diagnosing and intervening during critical incidents (Schäbe, 2019).

Ethical considerations often revolve around algorithmic decision-making in unavoidable accident scenarios, commonly framed as "trolley problems" (Gill, 2020) (Bergmann et al., 2018) (Bonnefon et al., 2016). XAI can provide transparency regarding the ethical principles encoded into an AV's behavior, explaining which values (e.g., protecting occupants versus external parties) were prioritized in each situation (Arkin, 2009) (Michael et al., 2020). This transparency is vital for public acceptance and for addressing potential societal concerns (Pettigrew et al., 2019). Moreover, XAI contributes to addressing biases that might be present in the AI models due to training data or algorithmic design, ensuring fair

outcomes across different demographics (Krontiris et al., 2020) (Leslie, 2019).

From a regulatory standpoint, XAI is becoming indispensable for certification, liability assignment, and legal compliance (2019) (Taeihagh & Lim, 2018). Regulators require mechanisms to verify that AVs adhere to safety standards and ethical guidelines (Sun, 2020). XAI provides the necessary audit trail for autonomous decisions, allowing authorities to attribute fault in the event of an accident (Iwan, 2019). It moves beyond mere compliance, enabling a framework for responsible innovation and ensuring public accountability (Leslie, 2019).

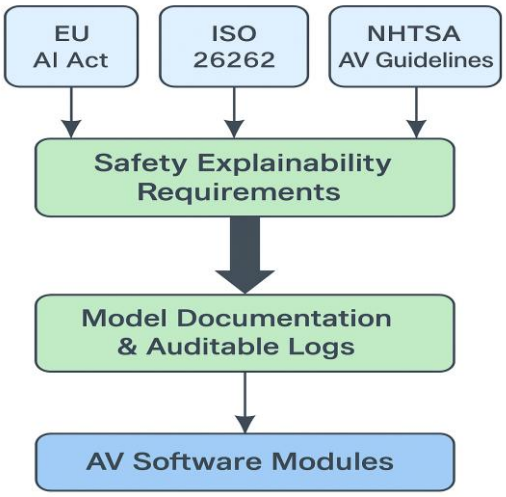

Figure 5: XAI and Regulatory Frameworks

Figure 5 shows how regulatory frameworks mandate various levels of explainability and auditability across AV subsystems to ensure accountability and liability transparency.

IV. ANALYSIS AND DISCUSSION

4.1 Technical and Design Challenges in Achieving Explainability

Achieving comprehensive explainability in autonomous vehicles faces significant technical and design challenges. The inherent complexity of deep learning models, which constitute the core of many AV systems, creates a fundamental hurdle (Padl, et al., 2020). These models often involve millions of parameters, making it difficult to trace a decision path or attribute causality to specific inputs. The black-box nature necessitates sophisticated post-hocXAI techniques, which themselves introduce computational overhead and may not fully capture the nuanced dynamics of the original model.

Real-time performance requirements of AVs further complicate XAI integration. Explanations must be generated rapidly, often within milliseconds, to be useful for real-time decision support or human intervention. This constraint limits the computational intensity of XAI methods that can be deployed on-board. Furthermore, the multi-modal nature of AV sensor data (e.g., cameras, lidar, radar) requires XAI techniques capable of explaining decisions derived from diverse and often conflicting information streams (Rosenfeld & Davis, 1986). Developing a unified explanation across these modalities remains a complex task. Another challenge involves the lack of standardized metrics for evaluating the quality of explanations, making it difficult to compare and benchmark different XAI approaches (Ahmad Fawad, 2023). The dynamic and continuously evolving nature of driving environments also means that explanations need to be robust and consistent across varied conditions, from clear weather to adverse scenarios, ensuring reliability and trustworthiness (Betz et al., 2019).

Another emerging challenge lies in adversarial robustness of explanations. Subtle input perturbations can significantly alter explanation outputs (e.g., heatmaps or feature attributes), leading to misleading insights. Moreover, explainability models may inadvertently propagate dataset or model biases—amplifying disparities in pedestrian detection or driving behavior across demographic or environmental conditions.

Integrating fairness-aware XAI techniques is essential to mitigate bias amplification in AV decision-making, particularly in scenarios involving vulnerable road users or underrepresented urban environments.

4.2 Implications of Explainability for Regulatory Compliance and Liability

The absence of adequate explainability in autonomous vehicle systems carries profound implications for regulatory compliance and liability

frameworks. Current legal systems are predicated on the ability to assign responsibility, which becomes problematic when an AI system's decision-making process is opaque (2019). In the event of an accident, XAI can provide crucial evidence by detailing the AV's perception, reasoning, and actions leading up to the incident (Vida & Váradi, 2018). This forensic capability supports investigations, helps determine fault, and informs potential liability assignments to manufacturers, software developers, or even fleet operators (Iwan, 2019) .

Regulatory bodies globally are grappling with establishing frameworks for AV deployment (Taeihagh & Lim, 2018) (Sun, 2020). Explainability is increasingly viewed as a prerequisite for certification and operational permits. It enables regulators to verify that AVs adhere to safety standards, ethical guidelines, and legal statutes, moving beyond mere performance validation to a deeper understanding of system behavior (Leslie, 2019). Without robust explanations, auditing AI systems for bias, fairness, and adherence to public policy objectives becomes exceedingly difficult (Krontiris et al., 2020). The current policy landscape often acknowledges these issues but requires more specific strategies for implementation (Taeihagh & Lim, 2018). The ability of XAI to provide a transparent audit trail of decisions is therefore not merely a technical nicety but a fundamental requirement for legal accountability and societal acceptance (Žolnerčíková, 2019).

4.3 Integration of XAI into Real-Time Decision-Making Architectures

Integrating XAI into the real-time decision-making architectures of autonomous vehicles presents a complex engineering challenge. AVs operate under strict latency requirements, demanding that perception, planning, and control modules execute their functions within milliseconds (Reid et al., 2019). The additional computational load imposed by XAI techniques, particularly post-hoc methods, must be carefully managed to avoid compromising the vehicle's responsiveness and safety (Ahmad Fawad, 2023).

One approach involves designing AV architectures that explicitly incorporate explanation generation as a parallel process, leveraging dedicated hardware or optimized algorithms. This could involve pre-computing certain explanations or employing highly efficient, model-specific XAI techniques. Another strategy is to differentiate the level of explainability based on the criticality of the decision. For instance, critical safety-related decisions might require detailed, probably correct explanations, while routine maneuvers could utilize simpler, high-level justifications. The system could dynamically adjust the depth and frequency of explanations based on the driving context and risk assessment. For human-in-the-loop scenarios, explanations must be concise and actionable, enabling timely human intervention (Padl, et al., 2020). This dynamic adaptation ensures that XAI enhances, rather than detracts from, the overall system performance and safety. The continuous monitoring of performance and the adaptive decision-making processes contribute to the refinement of these integrated systems (Ahmad Fawad, 2023).

4.4 Adversarial Robustness and Bias in AV Explainability

Adversarial attacks targeting perception and prediction modules can exploit vulnerabilities in deep neural networks, leading to unsafe planning or erroneous control actions in autonomous vehicles (Zhang et al., 2020) (Zheng et al., 2024) (Cao et al., 2022). Recent research demonstrates that adversarial robustness and explainability must be addressed jointly, as explainable models can help detect, localize, and mitigate such attacks in real time (Divya Bharat Mistry & Kaustubh Anilkumar Mandhane, 2024) (Nazat, Li, et al., 2024).

Recent studies demonstrate that explainability techniques can be leveraged to detect and localize adversarial attacks in real time, enhancing the overall safety of AV decision-making pipelines (Divya Bharat Mistry & Kaustubh Anilkumar Mandhane, 2024) (Yu et al., 2024). Furthermore, integrating adversarial robustness with XAI methods is increasingly recognized as essential for building resilient and trustworthy autonomous vehicle systems (Jiao et al., 2022) (Divya Bharat Mistry & Kaustubh Anilkumar Mandhane, 2024).

4.5 Emerging Solutions and Future Research Trajectories

Emerging solutions in XAI for autonomous vehicles are addressing current limitations and charting new research trajectories. One promising area is the development of inherently interpretable neural networks, which build transparency directly into the model architecture, aiming to bridge the gap between performance and interpretability (Grossberg, 2020). This contrasts with traditional black-box models requiring post-hoc analysis. Another solution involves causal XAI, which moves beyond correlation to identify direct cause-and-effect relationships in AV decision-making, offering more robust and trustworthy explanations (Ignatiev, 2020).

Future research trajectories will likely focus on:

- Standardized XAI Benchmarks: Establishing common datasets, metrics, and evaluation protocols to objectively compare and validate XAI methods specifically for AV applications (Ahmad Fawad, 2023).
- Adaptive Explainability: Developing systems that can tailor explanations based on the recipient (e.g., passenger, remote operator, regulator), context (e.g., routine driving, emergency), and cognitive load (Cunneen et al., 2019).
- Ethical Alignment: Further integrating ethical principles directly into AV decision-making algorithms and using XAI to explain how these principles are upheld during operation (Arkin, 2009).
- Human-AI Teaming: Exploring how XAI can facilitate seamless collaboration between human operators and autonomous systems, particularly in shared control scenarios or during handover processes (VINKHUYZEN & CEFKIN, 2016).
- Long-Term Explainability and Auditability: Designing systems capable of logging and explaining decisions over extended periods, providing a comprehensive audit trail for regulatory compliance and post-incident analysis.

These advancements will be critical for addressing the dynamic workload characteristics and the need for robust, ongoing model updates in AVs (Ahmad Fawad, 2023).

CONCLUSION

This study contributes to the field by (1) offering a structured taxonomy of XAI methods tailored for AV systems, (2) identifying key metrics for evaluating explanation quality, (3) analyzing trade-offs between interpretability and real-time performance, and (4) highlighting the regulatory and ethical implications of deploying XAI in critical safety systems.

5.1 Synthesis of Key Insights and Implications

Explainable Artificial Intelligence (XAI) is an indispensable component for the responsible development and deployment of autonomous vehicles. This review highlights that while AI models deliver unparalleled performance in AV functions, their inherent opacity presents significant challenges for safety, trust, and accountability. Key insights reveal a persistent trade-off between model interpretability and optimal performance, necessitating innovative hybrid architectures or inherently transparent AI designs. Human-machine interaction studies underscore the need for user-centric explanations that build appropriate trust and foster effective collaboration between humans and AVs.

The implications of robust XAI extend beyond technical functionality to crucial societal and regulatory domains. Explainability is fundamental for legal liability assignment in accident scenarios, allowing for forensic analysis of algorithmic decisions. It also provides a mechanism for regulators to audit AV systems for compliance with evolving safety standards and ethical guidelines. Without XAI, the path to widespread public and regulatory acceptance of autonomous vehicles remains fraught with obstacles, potentially hindering their societal benefits. The field demands continued focus on technical advancements that deliver real-time, context-aware explanations without compromising operational efficiency.

5.2 Recommendations and Prospective Pathways for XAI in Autonomous Vehicles

Based on the comprehensive review, several recommendations and prospective pathways for XAI in autonomous vehicles emerge:

1. Prioritize Inherently Interpretable Models: Research efforts should increasingly focus on developing AI models for AVs that are transparent by design, reducing reliance on post-hoc explanation techniques in safety-critical modules.
2. Develop Standardized XAI Metrics and Benchmarks: The establishment of industry-wide metrics and benchmarks for evaluating explanation quality, fidelity, and utility is crucial for comparative analysis and regulatory validation (Ahmad Fawad, 2023).
3. Invest in Adaptive and Context-Aware Explanations: Future XAI systems should be capable of tailoring explanations to the specific user and operational context, adjusting granularity and format dynamically to optimize comprehension and trust (Cunneen et al., 2019).
4. Integrate XAI Early in the Development Cycle:XAI considerations should be embedded from the initial design phase of AV software and hardware, rather than being an afterthought. This facilitates better integration and performance.
5. Foster Interdisciplinary Collaboration: Continued collaboration among AI researchers, automotive engineers, cognitive scientists, legal experts, and ethicists is vital to address the multifaceted challenges of XAI in AVs (Michael et al., 2020).

Prospective pathways include advancing causal inference methods for more robust explanations, exploring the use of distributed ledger technology for immutable audit trails of AI decisions (Padl, et al., 2020), and designing human-AI interfaces that intuitively convey machine intent and uncertainty. By embracing these pathways, the AV industry can build systems that are not only capable but also transparent, trustworthy, and accountable, accelerating their safe and ethical integration into society.

5.4 Future Research Directions and Emerging Trends

Future studies should prioritize the development of standardized benchmarks and evaluation metrics tailored specifically to XAI in autonomous vehicles to enable objective comparison and validation of explainability methods. Adaptive explainability that dynamically adjusts the level and modality of explanations based on user type, context, and cognitive load holds significant promise for enhancing human-machine interaction and trust. Additionally, integrating ethical considerations directly into AI decision-making frameworks, coupled with transparent explanations of these principles, is essential for societal acceptance and regulatory compliance. Research into long-term auditability, leveraging technologies such as distributed ledgers for immutable decision trails, will further underpin accountability and safety in AV deployment (Ahmad Fawad, 2023) (Padl, et al., 2020).

Emerging research highlights the integration of distributed ledger technology (DLT) as an immutable audit trail for AI model metadata and decision provenance in autonomous vehicles, enhancing both explainability and security (Padl, et al., 2020). Additionally, Trusted Execution Environments (TEEs) are being explored to securely manage and log inference processes, providing robust support for large-scale, data-intensive machine learning while preserving data confidentiality and supporting regulatory compliance (Padl, et al., 2020).

As AV technology becomes increasingly autonomous, the ability to explain, justify, and audit decisions in real time will be central to societal trust, legal accountability, and long-term deployment success.

REFERENCES

[1] Zöllner, J. M., & Schamm, T. (2015). Autonomous driving. In it - Information Technology (Vol. 57, Issue 4, pp. 213–214). Walter de Gruyter GmbH. https://doi.org/10.1515/itit-2015-0027

[2] Minaie, A., Sanati-Mehrizy, R., & Chambers, B. (2020). Autonomous Vehicles in Computer Engineering Program. In 2020 ASEE Virtual Annual Conference Content Access Proceedings. ASEE Conferences. https://doi.org/10.18260/1-2--34200

[3] Wylde, M. J. (2012.). Safe Motion Planning for Autonomous Driving. Wesleyan University. https://doi.org/10.14418/wes01.1.114

[4] Pandl, K. D., Thiebes, S., Schmidt Kraepelin, M., & Sunyaev, A. (2020). On the convergence

of artificial intelligence and distributed ledger technology: A scoping review and future research agenda (arXiv:2001.11017) [Preprint]. arXiv. https://arxiv.org/abs/2001.11017

[5] Betz, J., Heilmeier, A., Wischnewski, A., Stahl, T., & Lienkamp, M. (2019). Autonomous Driving—A Crash Explained in Detail. In Applied Sciences (Vol. 9, Issue 23, p. 5126). MDPI AG. https://doi.org/10.3390/app9235126

[6] Vida, G., & Váradi, P. (2018). Irregular operation of autonomous vehicles. In Production Engineering Archives (Vol. 20, Issue 20, pp. 8–11). Stowarzyszenie Menedzerow Jakosci i Produkcji. https://doi.org/10.30657/pea.2018.20.02

[7] Larasati, R., & Deliddo, A. (2020). Building a Trustworthy Explainable AI in Healthcare. Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from the INTERACT 2019 Workshops. https://doi.org/10.18573/book3.ab

[8] Hakimi, H. (2018). Trust Requirements Model for Developing Acceptable Autonomous Car. In Journal of Electrical and Electronic Engineering (Vol. 6, Issue 2, p. 59). Science Publishing Group. https://doi.org/10.11648/j.jeee.20180602.14

[9] Leslie, D. (2019). Understanding artificial intelligence ethics and safety. Social Science Research Network, abs/1906.05684. https://doi.org/10.5281/zenodo.3240529

[10] Krontiris, I., Grammenou, K., Terzidou, K., Zacharopoulou, M., Tsikintikou, M., Baladima, F., Sakellari, C., & Kaouras, K. (2020). Autonomous Vehicles: Data Protection and Ethical Considerations. In Computer Science in Cars Symposium (pp. 1–10). ACM. https://doi.org/10.1145/3385958.3430481

[11] Michael, K., Abbas, R., Roussos, G., Scornavacca, E., & Fosso-Wamba, S. (2020). Ethics in AI and Autonomous System Applications Design. In IEEE Transactions on Technology and Society (Vol. 1, Issue 3, pp. 114–127). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tts.2020.3019595

[12] Schäbe, H. (2019). Autonomous Driving – How to Apply Safety Principles. In Dependability (Vol. 19, Issue 3, pp. 21–33). Journal Dependability. https://doi.org/10.21683/1729-2646-2019-19-3-21-33

[13] Massoud, Y., & Laganiere, R. (2024). Learnable fusion mechanisms for multimodal object detection in autonomous vehicles. In IET Computer Vision (Vol. 18, Issue 4, pp. 499–511). Institution of Engineering and Technology (IET). https://doi.org/10.1049/cvi2.12259

[14] Pavel, M. I., Tan, S. Y., & Abdullah, A. (2022). Vision-Based Autonomous Vehicle Systems Based on Deep Learning: A Systematic Literature Review. In Applied Sciences (Vol. 12, Issue 14, p. 6831). MDPI AG. https://doi.org/10.3390/app12146831

[15] Patel, K., Beluch, W., Rambach, K., Cozma, A.-E., Pfeiffer, M., & Yang, B. (2021). Investigation of Uncertainty of Deep Learning-based Object Classification on Radar Spectra. In 2021 IEEE Radar Conference (RadarConf21) (pp. 1–6). IEEE.https://doi.org/10.1109/radarconf2147009.2021.9455269

[16] Nazat, S., Li, L., & Abdallah, M. (2024). XAI-ADS: An Explainable Artificial Intelligence Framework for Enhancing Anomaly Detection in Autonomous Driving Systems. In IEEE Access (Vol. 12, pp. 48583–48607). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/access.2024.3383431

[17] Teeti, I., Khan, S., Shahbaz, A., Bradley, A., & Cuzzolin, F. (2022). Vision-based Intention and Trajectory Prediction in Autonomous Vehicles: A Survey. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (pp. 5630–5637). International Joint Conferences on Artificial Intelligence Organization.https://doi.org/10.24963/ijcai.2022/785

[18] Limeros, S. C., Majchrowska, S., Johnander, J., Petersson, C., & Llorca, D. F. (2022). Towards Explainable Motion Prediction using Heterogeneous Graph Representations. Transportation Research Part C: Emerging Technologies,

abs/2212.03806.https://doi.org/10.48550/arXiv.2212.03806

[19] Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2022). Explanations in Autonomous Driving: A Survey. In IEEE Transactions on Intelligent Transportation Systems (Vol. 23, Issue 8, pp. 10142–10162). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tits.2021.3122865

[20] Gao, K., Yan, D., Yang, F., Xie, J., Liu, L., Du, R., & Xiong, N. (2019). Conditional Artificial Potential Field-Based Autonomous Vehicle Safety Control with Interference of Lane Changing in Mixed Traffic Scenario. In Sensors (Vol. 19, Issue 19, p. 4199). MDPI AG. https://doi.org/10.3390/s19194199

[21] Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. In Proceedings of the National Academy of Sciences (Vol. 116, Issue 50, pp. 24972–24978). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1820676116

[22] Ahmad Fawad, Et. al. (2023). Efficient Workload Allocation and Scheduling Strategies for AI-Intensive Tasks in Cloud Infrastructures. In Power System Technology (Vol. 47, Issue 4, pp. 82–102). Science Research Society. https://doi.org/10.52783/pst.160

[23] Pettigrew, S., Worrall, C., Talati, Z., Fritschi, L., & Norman, R. (2019). Dimensions of attitudes to autonomous vehicles. In Urban, Planning and Transport Research (Vol. 7, Issue 1, pp. 19–33). Informa UK Limited. https://doi.org/10.1080/21650020.2019.1604155

[24] Zhang, Q., Yang, X. J., & Robert, L. P. (2020). Expectations and Trust in Automated Vehicles. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–9). ACM. https://doi.org/10.1145/3334480.3382986

[25] Haspiel, J., Du, N., Meyerson, J., Robert Jr., L. P., Tilbury, D., Yang, X. J., & Pradhan, A. K. (2018). Explanations and Expectations. In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 119–120). ACM. https://doi.org/10.1145/3173386.3177057

[26] Žolnerčíková, V. (2019). Autonomous Vehicles and the Law: Technology, Algorithms and Ethics. Lim, Y. H. In Masaryk University Journal of Law and Technology (Vol. 13, Issue 2, pp. 415–421). Masaryk University Press. https://doi.org/10.5817/mujlt2019-2-12

[27] Anderson, J., Kalra, N., Stanley, K., Sorensen, P., Samaras, C., & Oluwatola, O. (2016). Autonomous Vehicle Technology: A Guide for Policymakers. RAND Corporation. https://doi.org/10.7249/rr443-2

[28] Sajjad, S., Abdullah, A., Arif, M., Faisal, M. U., Ashraf, M. D., & Ahmad, S. (2022). A Comparative Analysis of Camera, LiDAR and Fusion Based Deep Neural Networks for Vehicle Detection. In International Journal of Innovations in Science and Technology (Vol. 3, Issue 5, pp. 177–186). 50Sea. https://doi.org/10.33411/ijist/2021030514

[29] Wickramarachchi, R., Henson, C., & Sheth, A. (2024). Knowledge Graphs of Driving Scenes to Empower the Emerging Capabilities of Neurosymbolic AI. In IEEE Internet Computing (Vol. 28, Issue 6, pp. 62–67). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mic.2024.3494972

[30] Gyevnar, B., Droop, S., Quillien, T., Cohen, S. B., Bramley, N. R., Lucas, C. G., & Albrecht, S. V. (2025). People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior: Insights from Cognitive Science for Explainable AI. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1–18). ACM. https://doi.org/10.1145/3706598.3713509

[31] Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., & Zhu, J. (2023). Benchmarking Robustness of 3D Object Detection to Common Corruptions in Autonomous Driving. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1022–1032). IEEE. https://doi.org/10.1109/cvpr52729.2023.00105

[32] Wang, T., Kim, S., Ji, W., Xie, E., Ge, C., Chen, J., Li, Z., & Luo, P. (2023). DeepAccident: A

Motion and Accident Prediction Benchmark for V2X Autonomous Driving. AAAI Conference on Artificial Intelligence, 5599–5606. https://doi.org/10.48550/arXiv.2304.01168

[33] Wu, X., Wang, G., & Shen, N. (2023). Research on obstacle avoidance optimization and path planning of autonomous vehicles based on attention mechanism combined with multimodal information decision-making thoughts of robots. In Frontiers in Neurorobotics (Vol. 17). Frontiers MediaSA.https://doi.org/10.3389/fnbot.2023.1269447

[34] Li, X., Tseng, H. E., Girard, A., & Kolmanovsky, I. (2024). Autonomous Driving with Perception Uncertainties: Deep-Ensemble Based Adaptive Cruise Control. In 2024 IEEE 63rd Conference on Decision and Control (CDC) (pp. 8186–8192). IEEE. https://doi.org/10.1109/cdc56724.2024.10886150

[35] Grossberg, S. (2020). A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. In Frontiers in Neurorobotics (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fnbot.2020.00036

[36] Rokseth, B., Haugen, O. I., & Utne, I. B. (2019). Safety Verification for Autonomous Ships. In N. Karanikas, M. M. Chatzimichailidou, & M. Rejzek (Eds.), MATEC Web of Conferences (Vol. 273, p. 02002). EDP Sciences. https://doi.org/10.1051/matecconf/201927302002

[37] Kumari, D., & Bhat, S. (2021). Application of Artificial Intelligence in Tesla- A Case Study. In International Journal of Applied Engineering and Management Letters (pp. 205–218). Srinivas University.https://doi.org/10.47992/ijaeml.2581.7000.0113

[38] -, M. K., -, G. K., -, L. S., & -, A. T. (2024). Driving Towards Safety: The Role of ECUs and IMUs in Advanced Driver-Assistance Systems (ADAS). In International Journal For Multidisciplinary Research (Vol. 6, Issue 2). International Journal for Multidisciplinary Research (IJFMR). https://doi.org/10.36948/ijfmr.2024.v06i02.17022

[39] Cunneen, M., Mullins, M., Murphy, F., Shannon, D., Furxhi, I., & Ryan, C. (2019). Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics. In Cybernetics and Systems (Vol. 51, Issue 1, pp. 59–80). Informa UK Limited. https://doi.org/10.1080/01969722.2019.1660541

[40] VINKHUYZEN, E., & CEFKIN, M. (2016). Developing Socially Acceptable Autonomous Vehicles. In Ethnographic Praxis in Industry Conference Proceedings (Vol. 2016, Issue 1, pp. 522–534). Wiley. https://doi.org/10.1111/1559-8918.2016.01108

[41] Filiz, C. (2020). Can Autonomous Vehicles Prevent Traffic Accidents? In Accident Analysis and Prevention. IntechOpen. https://doi.org/10.5772/intechopen.93020

[42] Gill, T. (2020). How Important are Ethical Dilemmas to Potential Adopters of Autonomous Vehicles? Center for Open Science. https://doi.org/10.31234/osf.io/6vh8k

[43] Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making. In Frontiers in Behavioral Neuroscience (Vol. 12). Frontiers Media SA. https://doi.org/10.3389/fnbeh.2018.00031

[44] Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. In Science (Vol. 352, Issue 6293, pp. 1573–1576). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.aaf2654

[45] Arkin, R. C. (2009). An Ethical Basis for Autonomous System Deployment. Defense TechnicalInformationCenter.https://doi.org/10.21236/ada508646

[46] (2019). Product liability for autonomous vehicles. In Wiadomości Ubezpieczeniowe (Vol. 4, Issue 2019/4). The Polish Insurance Association. https://doi.org/10.33995/wu2019.4.1

[47] Taeihagh, A., & Lim, H. S. M. (2018). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. In Transport Reviews (Vol. 39, Issue 1, pp. 103–128). Informa UK Limited. https://doi.org/10.1080/01441647.2018.1494640

[48] Sun, Y. (2020). Construction of Legal System for Autonomous Vehicles*. In Proceedings of the 4th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2020). Atlantis Press. https://doi.org/10.2991/assehr.k.200316.131

[49] Iwan, D. (2019). Autonomous Vehicles – a New Challenge to Human Rights? In Przegląd Prawniczy Uniwersytetu im. Adama Mickiewicza (Vol. 9). Adam Mickiewicz University Poznan. https://doi.org/10.14746/ppuam.2019.9.04

[50] Rosenfeld, A., & Davis, L. S. (1986). Autonomous Vehicle Navigation. Defense Technical Information Center. https://doi.org/10.21236/ada170379

[51] Reid, T. G. R., Houts, S. E., Cammarata, R., Mills, G., Agarwal, S., Vora, A., & Pandey, G. (2019). Localization Requirements for Autonomous Vehicles. In SAE International Journal of Connected and Automated Vehicles (Vol. 2, Issue 3). SAE International. https://doi.org/10.4271/12-02-03-0012

[52] Zhang, Q., Hu, S., Sun, J., Chen, Q. A., & Mao, Z. M. (2022). On Adversarial Robustness of Trajectory Prediction for Autonomous Vehicles. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1513815147). IEEE. https://doi.org/10.1109/cvpr52688.2022.01473

[53] Zheng, J., Lin, C., Sun, J., Zhao, Z., Li, Q., & Shen, C. (2024). Physical 3D Adversarial Attacks against Monocular Depth Estimation in Autonomous Driving. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 24452–24461). IEEE. https://doi.org/10.1109/cvpr52733.2024.02308

[54] Cao, Y., Xiao, C., Anandkumar, A., Xu, D., & Pavone, M. (2022). AdvDO: Realistic Adversarial Attacks for Trajectory Prediction. European Conference on Computer Vision, 36–52. https://doi.org/10.48550/arXiv.2209.08744

[55] Divya Bharat Mistry, & Kaustubh Anilkumar Mandhane. (2024). Adversarial attacks and defense mechanisms for image classification deep learning models in autonomous driving systems. In International Journal of Science and Research Archive (Vol. 13, Issue 2, pp. 1998–1917). GSC OnlinePress.https://doi.org/10.30574/ijsra.2024.13.2.2328

[56] Nazat, S., Arreche, O., & Abdallah, M. (2024). On Evaluating Black-Box Explainable AI Methods for Enhancing Anomaly Detection in Autonomous Driving Systems. In Sensors (Vol. 24, Issue 11, p. 3515). MDPI AG. https://doi.org/10.3390/s24113515

[57] Yu, Z., Li, A., Wen, R., Chen, Y., & Zhang, N. (2024). PhySense: Defending Physically Realizable Attacks for Autonomous Systems via Consistency Reasoning. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (pp. 3853–3867). ACM. https://doi.org/10.1145/3658644.3690236

[58] Jiao, R., Liu, X., Sato, T., Chen, Q. A., & Zhu, Q. (2022). Semi-supervised Semantics-guided Adversarial Training for Trajectory Prediction. arXiv.Org, abs/2205.14230.https://doi.org/10.48550/arXiv.2205.14230

[59] Ignatiev, A. (2020). Towards Trustable Explainable AI. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (pp. 5154–5158). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/726