International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 322-326
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0044

A Conceptual Framework for Safety and Decision Justification in Autonomous Vehicles Using Explainable AI

P. Ramanjaneya Prasad[1], B. Durga Neelima[2], Shweta[3], A. Swapna[4]

[1]Assistant Professor, Dept. of Computer Science, Avanthi Degree & PG College, Hyderabad, Telangana, India

[2]Assistant Professor & Head, Dept. of Computer Science, Avanthi Degree & PG College, Hyderabad, Telangana, India

[3]Assistant Professor & Head, Dept. of Computer Science, Avanthi Degree & PG College, Hyderabad, Telangana, India

[4]Assistant Professor & Head, Dept. of Computer Science, Avanthi Degree & PG College, Hyderabad, Telangana, India

Emails: *karuprp@gmail.com[1], neeludurga1976@gmail.com[2], shwetapuranik84@gmail.com[3], swapnaaRroju22@gmail.com[4]*

Abstract

Autonomous Vehicles (AVs) have the potential to reduce road accidents and improve transportation efficiency significantly. However, safety concerns and the lack of transparency in decision-making remain major barriers to their widespread adoption. Modern AV systems rely heavily on complex Artificial Intelligence (AI) and Deep Learning models, which often function as black boxes, making it difficult to understand or justify their actions. This paper explores the critical role of safety mechanisms and decision justification in autonomous driving systems. We discuss the AV decision pipeline, identify safety challenges, and highlight the importance of Explainable Artificial Intelligence (XAI) techniques in improving trust, accountability, and regulatory compliance. The paper concludes by outlining open research challenges and future directions for safer and more transparent autonomous driving systems.

Keywords: *Autonomous Vehicles, Safety, Explainable AI, Decision Justification, Trustworthy AI*

1. Introduction

Autonomous vehicles have transformed modern transportation by enabling vehicles to sense their surroundings, make decisions, and navigate without human intervention [1], [2]. Rapid progress in sensors, machine learning techniques, and computational power has significantly accelerated the development of self-driving technologies [1]. Despite these advances, safety and transparency remain critical challenges, particularly in complex and uncertain real-world driving environments [2].

Many autonomous driving systems rely on deep neural networks for perception and decision-making tasks [3]. Although these models achieve high predictive accuracy, they often function as black boxes, making it difficult to understand why a vehicle chooses specific actions such as braking or lane changes [6]. In safety-critical domains like transportation, the ability to justify decisions is as important as accuracy itself [3]. Therefore, this work emphasizes the integration of safety mechanisms with decision justification to develop explainable and trustworthy autonomous vehicle systems

2. Autonomous Vehicle Decision Making Pipeline

An autonomous vehicle operates through a structured decision-making pipeline that includes perception, prediction, planning, and control stages [1]. Figure 1 shows Autonomous Vehicle Decision-Making Pipeline

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 322-326
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0044

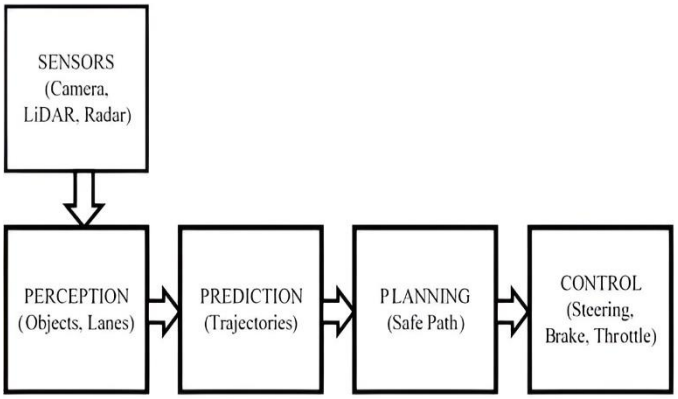

Figure 1 Autonomous Vehicle Decision-Making Pipeline

2.1 Perception

Perception is the initial stage of the autonomous driving pipeline and is responsible for understanding the surrounding environment. Autonomous vehicles employ multiple sensors, such as cameras, LiDAR, radar, ultrasonic sensors, and GPS, to collect raw environmental data [5], [12]. Each sensor has distinct advantages; cameras provide rich semantic information, LiDAR offers accurate depth measurements, and radar performs reliably in adverse weather conditions. Advanced deep learning models, including convolutional neural networks and transformer-based architectures, are used to detect and classify objects such as vehicles, pedestrians, cyclists, traffic signs, and lane markings [12]. Sensor fusion techniques combine data from multiple sensors to enhance robustness and reduce uncertainty, thereby improving perception reliability [5], [11]. Errors at this stage can propagate through the pipeline and directly compromise vehicle safety.

2.2 Prediction

The prediction module estimates the future behavior of surrounding road users based on their current states and historical motion patterns [1]. This includes predicting trajectories, speeds, and potential maneuvers of nearby vehicles and pedestrians. Since human behavior is inherently dynamic and context-dependent, prediction involves unavoidable uncertainty [2]. Probabilistic models, recurrent neural networks, long short-term memory networks, and graph-based methods are widely used to model interactions among multiple agents [2]. Accurate prediction enables proactive safety by allowing autonomous vehicles to anticipate dangerous scenarios such as sudden lane changes or unexpected pedestrian crossings [1].

2.3 Planning

During the planning stage, the autonomous vehicle determines an optimal course of action using outputs from perception and prediction modules. Generated trajectories must be safe, legally compliant, comfortable, and efficient [2]. Planning involves balancing multiple objectives, including collision avoidance, adherence to traffic rules, passenger comfort, and navigation efficiency. Rule-based, optimization-based, and learning-based planning algorithms are commonly applied [1]. Safety constraints such as maintaining safe distances and avoiding collisions are explicitly incorporated. In complex urban environments, planning must also consider ethical aspects, prioritizing human safety over operational efficiency [3], [4].

2.4 Control

The control module translates planned trajectories into executable vehicle commands such as steering angles, throttle inputs, and braking forces. Classical control techniques, including PID controllers and model predictive control, along with learning-based controllers, are used to ensure accurate actuation [1]. Control systems must operate in real time and remain stable under varying road and weather conditions. Any mismatch between planned trajectories and actual vehicle behaviour can introduce safety risks, making robust control strategies and continuous feedback mechanisms essential [1], [14].

3. Safety in Autonomous Vehicles

Safety in autonomous driving systems is achieved through a multi-layered, system-level approach spanning hardware, software, decision-making, and operational domains [14]. Each layer independently and collectively contributes to reducing risk and ensuring reliable vehicle behaviour in complex traffic environments. Figure 2 shows Flow Diagram of Safety Architecture in Autonomous Vehicles

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 322-326
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0044

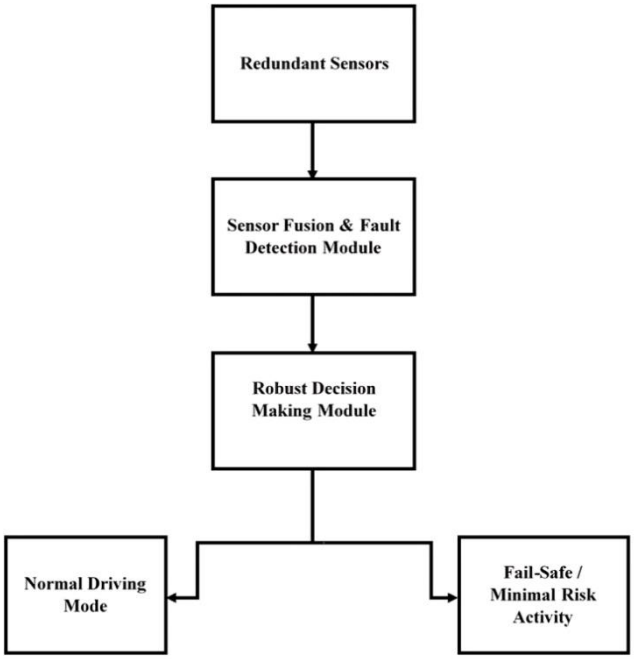

Figure 2 Flow Diagram of Safety Architecture in Autonomous Vehicles

3.1 Sensor and System Redundancy

Redundancy is a fundamental safety principle in autonomous vehicle design. Critical components such as sensors, computing units, communication links, and power supplies are duplicated to avoid single-point failures [14]. If one sensor fails or produces unreliable data, alternative sources can maintain situational awareness. Redundant architectures improve fault tolerance and allow vehicles to continue operating safely or transition to minimal-risk states, thereby enhancing system reliability and compliance with functional safety requirements [10].

3.2 Robust Decision-Making

Robust decision-making ensures safe vehicle behaviour under uncertainty, noise, and unexpected environmental conditions [1]. Decision algorithms are designed to behave conservatively in ambiguous situations, prioritizing safety over performance metrics such as speed or travel time. Risk-aware planning, uncertainty modelling, and worst-case scenario analysis are commonly employed to handle edge cases. Adaptive behaviour and continuous system monitoring further enable vehicles to adjust decisions in response to changing traffic conditions [2].

3.3 Runtime Monitoring & Safety Assurance

Runtime monitoring continuously evaluates system health and operational behaviour during vehicle operation [15]. This includes monitoring sensor reliability, model confidence levels, and system latency [8], [9]. When anomalies or confidence degradation are detected, appropriate safety mechanisms are triggered. Runtime verification is particularly important for detecting rare or unseen scenarios that may not have been covered during training or testing, enabling early corrective actions [15].

3.4 Fail-Safe and Minimal-Risk

Fail-safe mechanisms are activated when safe autonomous operation cannot be guaranteed. These include controlled braking, safe stopping, or pulling over in response to severe sensor failures or system malfunctions [14]. Minimal-risk activities aim to reduce harm to passengers, pedestrians, and surrounding vehicles and are essential for regulatory approval and real-world deployment of autonomous vehicles [7].

4. Decision justification and Explainable AI

Decision Justification refers to the ability of an autonomous vehicle to explain why a specific action was taken

4.1 Need for Explainability

Explainability is essential for building trust in autonomous vehicle systems among passengers, engineers, regulators, and legal authorities [3], [13]. Without transparency, it becomes difficult to validate system behaviour or assign responsibility in the event of an accident. Explainability also supports system debugging and improvement by revealing hidden biases, incorrect assumptions, or data limitations. In safety-critical systems, explainable decision-making is a fundamental requirement rather than an optional feature [6].

4.2 Explainable AI Techniques

Explainable AI techniques provide insights into the internal reasoning of autonomous vehicle models. Feature attribution methods such as SHAP and LIME identify the most influential input features affecting a model's decision [16]. Attention-based models highlight critical regions in sensor data, such as pedestrians or traffic signs. Hybrid and rule-based explanation approaches combine human-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 322-326
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0044

understandable logic with learning-based performance, enabling both real-time and post-hoc explanations of system behaviour [3], [13]. Figure 3 shows Methodology of XAI

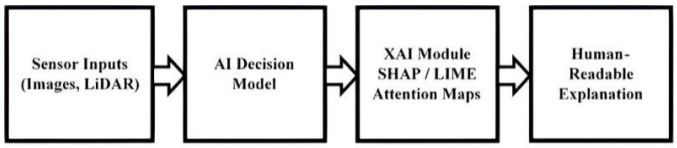

Figure 3 Methodology of XAI

4.3 Benefits of Decision Justification

Decision justification improves trust, accountability, and regulatory compliance in autonomous driving systems [13]. Clear explanations assist developers in validating system behaviour and help investigators analyse incidents and determine liability. Overall, explainable and justified decisions are essential for deploying autonomous vehicles as trustworthy participants in real-world transportation systems [3].

5. Challenges and Research Gaps

Despite advancements, a number of challenges still exist:

- Real-time explainability without losing performance
- Finding a balance between interpretability and model accuracy
- Standardization of formats for explanations
- Managing uncommon and unobserved driving situations
- Implementing AVs in real-world settings requires filling in these gaps.

6. Future Directions

- Future research to concentrate on:
- Combining XAI and causal reasoning to make safer choices
- Creating explanations that are human centered for various stakeholders
- Creating legal frameworks for autonomous systems that can be explained
- Integrating multi-modal explanations from control data, sensors, and vision

Conclusion

For autonomous vehicles to function in environments that are real, safety and decision reasoning are essential requirements. Even though AI-driven adoption and trust are limited by their lack of transparency. By making autonomous decisions transparent, verifiable, and accountable, explainable AI offers an achievable way to close this gap. The effective implementation of autonomous cars in upcoming intelligent transportation systems will depend on ensuring both safety and explainability.

Acknowledgements

The authors would like to express their sincere gratitude to the Management of the College and the Department for their constant support and encouragement throughout this research work. The continuous guidance, academic freedom, and research-oriented environment provided by the institution have greatly contributed to the successful completion of this study. We also acknowledge the motivation and encouragement extended by the management and department to pursue quality research and to actively publish scholarly work, which inspires us to contribute more significantly to research in the future.

References

[1]. Cao, Y., Chen, Y., Liu, L., & Zhu, J. (2022). Research prospect of autonomous driving decision technology under complex traffic scenarios. MATEC Web of Conferences, https://doi.org/10.1051/matecconf/202235503031

[2]. Liu, Y., & Diao, S. (2024). An automatic driving trajectory planning approach in complex traffic scenarios based on integrated driver style inference and deep reinforcement learning. PloS One, 19(1), e0297192. https://doi.org/10.1371/journal.pone.0297192

[3]. Atakishiyev, S., Yao, H., Salameh, M., & Goebel, R. (2021). Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. Cornell University. https://doi.org/10.48550/arxiv.2112.11561

[4]. Atakishiyev, S., Salameh, M., & Goebel, R. (2024). Safety Implications of Explainable Artificial Intelligence in End-to-End Autonomous Driving. https://doi.org/10.48550/arxiv.2403.12176

[5]. Dai, Z., Guan, Z., Chen, Q., Sun, F., & Xu, Y. (2024). Enhanced Object Detection in Autonomous Vehicles through LiDAR—Camera Sensor Fusion. World Electric Vehicle Journal, 15(7), 297. https://doi.org/10.3390/wevj15070297

[6]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Cornell University. https://doi.org/10.48550/arxiv.1702.08608

[7]. Antona-Makoshi, J., Terranova, P., Hatchett, A., Kefauver, K., Sullivan, K., Ali, G., & Williams, V. (2025). Leveraging the Automated Mobility Partnership (AMP) to Support the Evaluation of Safety of the Intended Functionality (SOTIF) in Automated Driving Systems. 1. https://doi.org/10.4271/2025-01-8674

[8]. Colin Paterson, Richard Hawkins, Chiara Picardi, Yan Jia, Radu Calinescu, Ibrahim Habli. (2025). Safety assurance of Machine Learning for autonomous systems, https://doi.org/10.1016/j.ress.2025.111311

[9]. Patel, M., Jung, R., and Khatun, M., "A Systematic Literature Review on Safety of the Intended Functionality for Automated Driving Systems," SAE Technical Paper 2025-01-5030, 2025, doi:10.4271/2025-01-5030.

[10]. Neto, A. V. S., Camargo, J. B., Cugnasca, P. S., & Almeida, J. R. (2022). Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review on the State of the Art and Guidelines for Future Work. IEEE Access, 10, 130733–130770. https://doi.org/10.1109/access.2022.3229233

[11]. Abdullah, H., Ali, M., Naqvi, S., Khan, I., Faiz, A., & Majid, A. (2025). Multimodal Sensor Fusion in Autonomous Driving: A Deep Learning-Based Visual Perception Framework. Kashf Journal of Multidisciplinary Research, 2(06), 100–123. https://doi.org/10.71146/kjmr490

[12]. Liu, Haibin & Wu, Chao & Wang, Huanjie. (2023). Real time object detection using LiDAR and camera fusion for autonomous driving. Scientific Reports. 13. 10.1038/s41598-023-35170-z.

[13]. Kuznietsov, A., Albrecht, S. V., Wang, C., Peters, S., & Gyevnar, B. (2024). Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review. IEEE Transactions on Intelligent Transportation Systems, 25(12), 19342–19364. https://doi.org/10.1109/tits.2024.3474469

[14]. Matos, F., Bernardino, J., Durães, J., & Cunha, J. (2024). A Survey on Sensor Failures in Autonomous Vehicles: Challenges and Solutions. Sensors (Basel, Switzerland), 24(16), 5108. https://doi.org/10.3390/s24165108

[15]. Stemmer, R., Saxena, I., Panneke, L., Grundt, D., Austel, A., Möhlmann, E., & Westphal, B. (2025). Runtime monitoring of complex scenario-based requirements for autonomous driving functions. Science of Computer Programming, 244, https://doi.org/10.1016/j.scico.2025.103301

[16]. Tahir, H. A., Alayed, W., Hassan, W. U., & Haider, A. (2024). A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration. Sensors, 24(21), 6776. https://doi.org/10.3390/s24216776