International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

Explainable Artificial Intelligence (Xai) In Self-Driving Cars: A Review Of Techniques For Transparent Perception, Prediction and Decision-Making

Kohakade Vaishnavi Sanjay[1]; Prof. M. C. Kolse[2]; Prof. S. K. Said[3]

Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Self-driving cars use cameras, radar, LiDAR and deep learning to detect objects, recognize traffic signs, identify pedestrians and make driving decisions without direct human control. Most of these models behave as black boxes, making it difficult to understand why a vehicle braked, changed lanes or failed to detect an obstacle. This lack of transparency creates problems for safety assurance, debugging, user trust, accountability and regulation. Explainable Artificial Intelligence (XAI) addresses these problems by providing understandable reasons for AI decisions. This paper reviews XAI for autonomous driving across the perception, prediction and planning stages. It examines post-hoc methods such as LIME, SHAP and gradient-based saliency maps, visual attention and VisualBackProp for end-to-end steering models, textual explanations that describe driving actions in natural language, object-induced action explanations, and inherently interpretable designs. Stakeholder needs of developers, passengers, safety assessors and regulators are analysed. A layered XAI-enabled autonomous driving architecture is proposed, and research gaps, objectives and a comparison of explanation techniques are presented. The review concludes that integrating explainability throughout the driving stack, and evaluating explanations for faithfulness and usefulness, is essential for safe, reliable and trustworthy autonomous transportation.

Index Terms—Explainable artificial intelligence, self-driving cars, autonomous vehicles, interpretability, saliency maps, LIME, SHAP, road safety

I. Introduction

Autonomous vehicles promise fewer accidents caused by human error, greater mobility for people who cannot drive and more efficient traffic. Modern self-driving systems rely on deep neural networks to interpret sensor data and choose actions. These networks are accurate in many conditions, but they are opaque: even their designers cannot easily state why a particular decision was made.Opacity is a serious problem in a safety-critical domain. When an autonomous car makes an unexpected manoeuvre, engineers need to diagnose the cause, passengers want reassurance, insurers and courts need to assign responsibility, and regulators need evidence that the system is safe. Surveys of XAI [1], [2] identify such high-stakes settings as the primary motivation for explainability.

This paper reviews XAI techniques applied to self-driving cars, organized by the stages of the driving pipeline, and proposes an architecture that makes explainability a built-in feature rather than an afterthought.

II. Theoretical Background

2.1 Levels of Driving Automation

SAE J3016 [3] defines six levels of automation, from Level 0 with no automation to Level 5 with full automation in all conditions. At Levels 3 and above, the system rather than the driver monitors the environment, which greatly increases the need for transparent behaviour.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

2.2 The Autonomous Driving Pipeline

A modular stack consists of perception (detecting and tracking objects, lanes and signs), prediction (forecasting the motion of other road users), planning (choosing a path and manoeuvre) and control (steering, throttle and braking). End-to-end systems learn a direct mapping from sensor input to control [4]. Yurtsever *et al.* [5] review both approaches.

2.3 Categories of Explanation

Fig. 1 classifies XAI techniques. Post-hoc methods explain an already trained model; intrinsic methods use models that are interpretable by design. Explanations may be local, for one decision, or global, describing overall behaviour.

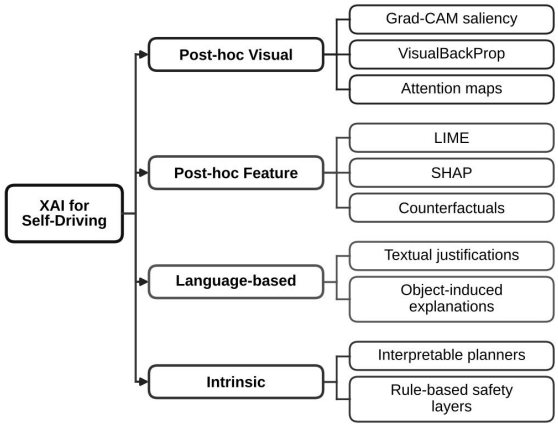


Fig. 1. Classification of XAI techniques for autonomous driving.

III. Survey of Existing Approaches

Bojarski *et al.* [4] trained a CNN to steer a car directly from camera images, demonstrating end-to-end driving. To understand what the network learned, the same group proposed VisualBackProp [6], an efficient method that highlights the image regions contributing most to the steering output; the highlighted regions corresponded to lane markings and road edges, increasing confidence in the model.

Kim and Canny [7] introduced a visual attention model for steering and refined the attention maps with causal filtering to keep only regions that actually affect the output. Kim *et al.* [8] went further by generating textual explanations such as "the car slows down because the light is red," trained on the BDD-X dataset of driving videos annotated with descriptions and justifications. Xu *et al.* [9] proposed predicting actions together with explanations that refer to the specific objects causing them, such as a pedestrian or a stopped vehicle.

General-purpose methods are also applied. LIME [10] explains a single prediction by fitting a simple local model around it, SHAP [11] attributes predictions to input features using Shapley values, and Grad-CAM [12] produces class-specific heat maps for CNN-based perception.

Zablocki *et al.* [13] reviewed explainability for vision-based driving and stressed that explanations must be faithful to the model and useful to the intended audience. Atakishiyev *et al.* [14] surveyed XAI for autonomous driving and proposed a

framework linking explanations to safety and regulation. Doshi-Velez and Kim [15] argued for rigorous evaluation of interpretability rather than relying on intuition. Table I summarizes the reviewed literature.

Ref.	Method	Key Findings	Limitations
[4]	End-to-end CNN steering	Drove from camera images directly	Black-box decisions
[6]	VisualBackProp	Fast visualization of influential regions	Steering task only
[7]	Causal visual attention	Attention filtered to causal regions	Attention not always faithful
[8]	Textual explanations (BDD-X)	Natural-language action justifications	Needs annotated videos
[9]	Object-induced action explanations	Explains actions via causal objects	Limited action set
[10]	LIME local surrogates	Model-agnostic local explanations	Unstable; slow for real time
[12]	Grad-CAM	Class-specific CNN heat maps	Coarse resolution
[13]	Review of explainability in vision-based driving	Defined requirements and challenges	Survey; no new method
[14]	Comprehensive XAI survey for AVs	Framework linking XAI to safety and regulation	Conceptual framework

TABLE I Summary of Literature Survey

IV. Comparative Study

Table II compares explanation techniques. Saliency methods are fast and visual but can be noisy; LIME and SHAP are model-agnostic but expensive; attention is cheap but not always faithful; and textual explanations are the most understandable for passengers but require annotated training data.

Technique	Scope	Runtime	Main Audience
Grad-CAM / saliency	Local, visual	Fast	Engineers
Attention maps	Local, visual	Very fast	Engineers
LIME	Local	Slow	Engineers, analysts
SHAP	Local + global	Slow–Medium	Engineers, regulators
Counterfactuals	Local	Medium	Investigators
Textual explanations	Local	Fast	Passengers

TABLE II Comparison of Explanation Techniques

V. Research Gaps

The following gaps are identified:

- 1. Most XAI studies address perception or end-to-end steering; explanations for prediction and planning modules are scarce.
- 2. Faithfulness of explanations, that is, whether they truly reflect the model's reasoning, is rarely measured.
- 3. Explanations are seldom tailored to different stakeholders such as passengers, engineers and regulators.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

- 4. Real-time generation of explanations within vehicle computing limits is not well studied.
- 5) Links between explanations and safety standards such as ISO 21448 [16] are still immature.

VI. Aims and Objectives

The objectives of this review are:

- 1. To study the need for explainability in autonomous driving.
- 2. To review XAI techniques for perception, prediction and decision-making.
- 3. To propose an XAI-enabled autonomous driving architecture.
- 4. To compare explanation techniques by scope, fidelity and runtime cost.
- 5. To identify challenges and future research directions.

VII. Proposed Architecture

Fig. 2 shows the proposed architecture. Each driving module produces both its normal output and an explanation, which are combined by an explanation manager and delivered to different stakeholders.

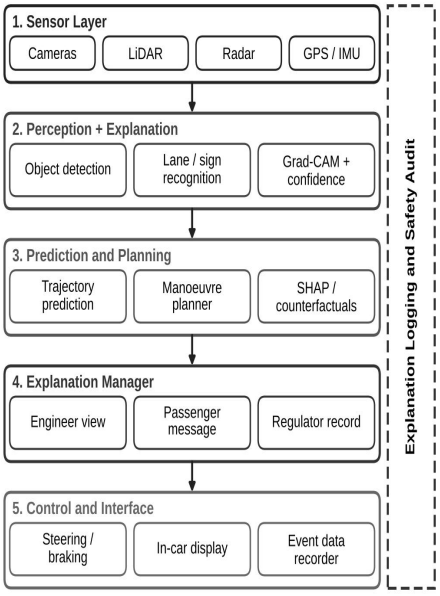


Fig. 2. Proposed XAI-enabled autonomous driving architecture.

7.1 Sensor Layer

Cameras, radar, LiDAR, GPS and inertial sensors provide raw data, which is time-synchronized and calibrated.

7.2 Perception and Explanation Layer

Detection and segmentation networks identify vehicles, pedestrians, lanes and signs. Grad-CAM heat maps and detection confidence scores explain what was seen and how certain the system is.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

7.3 Prediction and Planning Layer

Trajectory predictors forecast the movement of nearby road users, and a planner selects a manoeuvre. SHAP values or attention weights indicate which objects and features most influenced the decision, and counterfactual explanations [17] state what change would have led to a different action.

7.4 Explanation Manager

This module converts technical explanations into audience-appropriate forms: detailed visualizations and logs for engineers, concise messages such as “braking for pedestrian crossing” for passengers, and structured records for investigators.

7.5 Interface and Logging Layer

Explanations are shown on the in-vehicle display and stored in an event data recorder, supporting post-incident analysis and regulatory audits.

VIII. Process Flow and Evaluation

The workflow is shown in Fig. 3. For a perception CNN, Grad-CAM computes the importance weight of feature map A^k for class c as

$$\alpha_k^c = (1/Z)\, \Sigma_i \Sigma_j\, \partial y^c / \partial A^k_{ij} \quad (1)$$

and the heat map is the rectified weighted sum of feature maps. For tabular decision features, SHAP assigns each feature a contribution φ_i such that the prediction equals the baseline plus the sum of contributions:

$$f(x) = \varphi_0 + \Sigma_i\, \varphi_i \quad (2)$$

Explanation fidelity can be tested by deletion: removing the regions or features marked most important should change the output more than removing random ones.

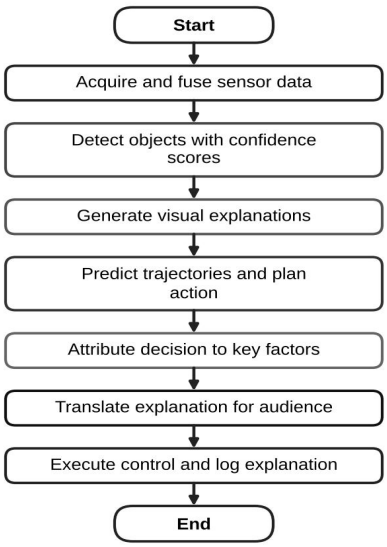


Fig. 3. Decision and explanation workflow in an XAI-enabled vehicle.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

IX. Limitations and Future Scope

Table III lists major challenges. Future directions include explanation-aware training that makes models inherently more interpretable, natural-language explanations generated by vision–language models, standardized benchmarks for faithfulness, and regulatory guidelines that specify what explanations autonomous vehicles must provide. Adversarial examples on traffic signs [18] show that explanations can also help detect when a model is being fooled.

Challenge	Mitigation
Faithfulness	Deletion / insertion tests; causal methods
Real-time limits	Lightweight methods; on-demand detail
Different audiences	Explanation manager with tailored views
Adversarial inputs	Explanations to flag abnormal attention
Regulation	Explanation logs aligned with safety standards

TABLE III Challenges and Mitigation

X. Conclusion

Explainable AI is a key enabler for the safe deployment of self-driving cars. The reviewed literature shows that saliency maps, attention, textual and object-level explanations, and feature-attribution methods can reveal why autonomous systems perceive and act as they do, helping developers find errors, passengers build trust and regulators assess safety. Explanations must be faithful, timely and adapted to their audience. Integrating XAI across perception, prediction and planning will support safer, more reliable and more trustworthy autonomous transportation.

10.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

References

1. A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.
2. A. Barredo Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Inf. Fusion, vol. 58, pp. 82–115, 2020.
3. SAE International, “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles,” SAE Standard J3016, 2021.
4. M. Bojarski et al., “End to end learning for self-driving cars,” arXiv:1604.07316, 2016.
5. E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, “A survey of autonomous driving: Common practices and emerging technologies,” IEEE Access, vol. 8, pp. 58443–58469, 2020.
6. M. Bojarski et al., “VisualBackProp: Efficient visualization of CNNs for autonomous driving,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2018, pp. 4701–4708.

7. J. Kim and J. Canny, “Interpretable learning for self-driving cars by visualizing causal attention,” in Proc. IEEE ICCV, 2017, pp. 2942–2950.
8. J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, “Textual explanations for self-driving vehicles,” in Proc. ECCV, 2018, pp. 563–578.
9. Y. Xu et al., “Explainable object-induced action decision for autonomous vehicles,” in Proc. IEEE/CVF CVPR, 2020, pp. 9523–9532.
10. M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?”: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
11. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS, vol. 30, 2017.
12. R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE ICCV, 2017, pp. 618–626.
13. É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, “Explainability of deep vision-based autonomous driving systems: Review and challenges,” Int. J. Comput. Vis., vol. 130, no. 10, pp. 2425–2452, 2022.
14. S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions,” IEEE Access, vol. 12, 2024.
15. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv:1702.08608, 2017.
16. International Organization for Standardization, “Road vehicles—Safety of the intended functionality,” ISO 21448:2022, 2022.
17. S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” Harvard J. Law Technol., vol. 31, no. 2, pp. 841–887, 2018.
18. K. Eykholt et al., “Robust physical-world attacks on deep learning visual classification,” in Proc. IEEE/CVF CVPR, 2018, pp. 1625–1634.