Narrative Review Article

A Review of AI Safety and Trustworthiness in Autonomous Vehicles

Elston Su

Stony Point High School, 10010 De Soto Ave, Chatsworth, CA 91311, United States

ABSTRACT

This narrative review examines how autonomous-vehicle safety depends on the combined roles of perception, robustness, explainability, and ethical alignment. It argues that reliable perception and robust model behavior form the technical foundation of safe decision-making, while explainability enables transparency, accountability, and system-level oversight. In addition, ethical considerations, including fairness, responsibility, and bias mitigation, are shown to be inseparable from safety in AI-driven driving systems. This article reviews recent research, comparing key technical challenges, robustness limitations, and interpretability requirements across autonomous driving architectures. The review identifies major gaps in existing work, including limited robustness under rare and unpredictable scenarios, insufficient dataset diversity, and the absence of unified safety and certification frameworks. It concludes that both industry and regulatory bodies must adopt stronger certification practices and responsible deployment strategies to ensure that autonomous-vehicle systems remain trustworthy, transparent, and safe.

Keywords: Autonomous vehicles; cybersecurity; explainable AI (XAI); machine learning safety; perception robustness; safety assurance; trustworthiness

INTRODUCTION

In recent years, modern vehicle companies have begun implementing autonomous driving technologies in their vehicles. For example, Tesla, an electric vehicle manufacturer, is widely known for its autonomous driving features. In technology-focused cities such as Austin or San Francisco, fully autonomous driving taxis can already be observed in operation. At the core of all autonomous driving systems is artificial intelligence (AI). AI is not merely an add-on to autonomous vehicles; rather, it is the central engine that enables vehicles to perceive their environment, make decisions, and operate in complex, dynamic conditions. Without AI, the development of highly autonomous vehicles would not be possible.

Despite these potential benefits, significant safety concerns accompany the use of AI in autonomous driving. As highlighted in recent literature, machine learning, particularly through artificial neural networks (ANNs), forms the backbone of modern AI systems. Because many AI systems rely on ANNs, traditional safety procedures are often inadequate, requiring the development of new safety approaches. Traditional engineering safety methods were designed for systems whose behavior can be explicitly specified, understood, and hierarchically decomposed. AI-based systems, however, violate many of these assumptions, which is why conventional safety methods do not transfer directly.

AI systems often operate as black-box neural

Corresponding author: Elston Su, E-mail: elstonsu.17@gmail.com.

Accepted April 23, 2026
https://doi.org/10.70251/HYJR2348.42499503

networks, with internal processes that cannot be easily interpreted in human-readable terms. This lack of interpretability prevents engineers from fully understanding why specific decisions are made or from identifying hidden failure modes, thereby undermining core safety activities. Verification and validation are also significantly more challenging because AI behavior emerges from training data rather than explicit rules, and autonomous driving involves open-ended environments that cannot be exhaustively tested. As discussed by Wäschle *et al*. (1), existing safety standards such as ISO 26262 assume deterministic and traceable software architectures, assumptions that neural network-based systems frequently fail to meet.

In autonomous driving, safety concerns extend beyond purely technical failures to include issues of fairness, ethics, and societal impact. Ethical questions arise regarding how an AI system should prioritize safety in critical situations, such as whether to protect vehicle occupants at the expense of pedestrians or vice versa. These dilemmas highlight the difficulty of encoding moral decision-making into automated systems and determining what constitutes the "safest" or most acceptable outcome. Furthermore, questions of accountability remain unresolved: if an AI-driven vehicle causes an accident, responsibility may lie with the driver, the manufacturer, or the system designers (2). These challenges are not only technical but also societal, as they influence public trust and regulatory decisions. The objective of this narrative review is to bridge the gap between technical AI performance and human-centric trustworthiness. Specifically, this review seeks to answer: 1) How do limitations in AI perception and data diversity impact vehicle safety? 2) To what extent does explainability bridge the gap between black-box models and regulatory accountability? and 3) What unified frameworks are necessary to align ethical decision-making with technical performance?

TECHNICAL FOUNDATIONS AND SAFETY CHALLENGES OF AI IN AUTONOMOUS DRIVING

AI is integrated into autonomous driving systems through two main architectural approaches, each with distinct safety implications. In the modular pipeline design, perception, prediction, planning, and control are separated into components, allowing engineers to inspect intermediate outputs. In contrast, end-to-end learning approaches map raw sensor inputs directly to driving actions using a single neural network, reducing hand-crafted components but creating a black-box system with limited transparency. While modular pipelines support clearer accountability, end-to-end models introduce challenges because their internal decision processes are difficult to interpret.

Ensuring the safety of autonomous vehicles relies on formal safety processes such as verification, validation, and testing. These methods have been effective for traditional software systems; however, they become significantly more difficult to apply when AI is involved. Unlike conventional code, AI systems learn their behavior from data, which makes it difficult for engineers to predict system behavior in every possible situation (3). This uncertainty complicates validation, as it is challenging to guarantee that the system will behave safely under all real-world conditions. Testing is also limited, since no dataset can fully represent the vast range of scenarios encountered in real traffic.

Furthermore, existing safety standards such as ISO 26262 were developed for deterministic, rule-based software and do not adequately address the risks introduced by neural networks (1). Training datasets are inherently incomplete because they cannot capture every possible road scenario or rare event (1, 3). This limitation becomes critical when the vehicle operates in unfamiliar environments, potentially leading to unreliable predictions. Additionally, AI models often exhibit non-linear behavior, where small changes in lighting or sensor noise can result in unexpected changes in system output.(25) These factors increase safety risks by making system behavior less predictable in edge cases.

ADVANCES IN PERCEPTION, ROBUSTNESS, AND SECURITY FOR AUTONOMOUS DRIVING

Perception systems form the foundation of autonomous driving. Autonomous vehicles rely on sensors such as cameras, LiDAR, and radar to detect vehicles, pedestrians, and traffic signs. If these objects are incorrectly perceived, the entire decision-making process becomes unsafe (4). Misclassification or incorrect estimation of speed can result in delayed braking or unsafe steering. Three-dimensional object detection methods generally fall into three main categories: camera-based approaches, LiDAR-based methods, and multi-modal approaches which combine data from multiple sensors to achieve improved performance (4).

A significant portion of recent literature focuses on

the vulnerability of perception systems to adversarial perturbations. Research has demonstrated that deep neural networks can be deceived by "physical-world" attacks, such as strategically placed tape on traffic signs that cause an AI to misclassify a 'Stop' sign as a 'Speed Limit' sign (5, 6). To combat this, researchers are exploring adversarial training and formal verification methods. These techniques aim to provide mathematical guarantees that a model's output will remain constant despite sensor noise or intentional tampering (7-10).

Cybersecurity threats pose serious risks because malicious attacks can directly influence AI behavior (15). Adversarial attacks may manipulate sensor inputs to cause misclassification, while data poisoning can corrupt training datasets (5, 15). These threats highlight the importance of securing AI systems against both digital and physical attacks. Extensive testing and continuous post-deployment monitoring are essential to detect unexpected behaviors and update systems as threats evolve. (3, 16, 17, 24)

TRUSTWORTHINESS, ETHICS, AND SYSTEM-LEVEL INTERACTIONS

Trust is a key factor for public acceptance and regulatory approval. Transparency plays a central role in building trust by making AI-driven decisions more understandable. Explainable artificial intelligence (XAI) refers to techniques designed to make AI processes understandable to humans. XAI enables insight into how AI models arrive at conclusions rather than treating them as opaque boxes (11, 12). By clarifying the reasoning behind AI decisions, XAI enhances transparency and supports safer vehicle systems.

Different stakeholders require different types of explanations because they evaluate the technology in distinct ways. Explanations should be tailored to the user's level of expertise, whether they are intended for passengers, engineers, or regulators (13). Fairness and bias directly influence how autonomous vehicles behave in morally complex situations, as incomplete training data can cause inconsistent performance across different populations (2, 11). Ethical shortcomings ultimately undermine safety by introducing unpredictability in high-stakes situations where reliability is most critical (2, 11).

The literature demonstrates that safety, robustness, perception, and explainability are deeply interconnected (1, 11, 12, 15). Perception represents the first link in this chain, as errors in sensor detection can immediately undermine safety. Explainability connects these components by making failures visible and traceable (11, 12). Safe autonomous driving can only be achieved when perception, robustness, and transparent decision-making are addressed as an integrated whole (1, 11, 15).

Trade-offs between performance and interpretability are essential in system design. For example, "End-to-End" models often achieve high driving smoothness but provide no internal traceability. In contrast, modular systems allow an engineer to see exactly which module failed (e.g., the detection module vs. the planning module), providing the transparency required for certification at the potential cost of system latency. Beyond technical robustness, the shift toward "Safety of the Intended Functionality" (SOTIF), or ISO 21448, addresses hazardous behaviors that occur without a specific system failure, such as a vehicle misinterpreting a reflection on a wet road (14) (Table 1).

Table 1. *Summary of key challenges, safety implications, and future directions for AI safety and trustworthiness in autonomous vehicles.*

Core Pillar	Technical Challenge	Safety Implication	Future Direction
Perception	Sensor noise, occlusion, and lighting variability	Misclassification of obstacles or delayed detection	Multi-modal sensor fusion and edge-case datasets
Robustness	Distribution shift from training data (edge cases)	Unpredictable behavior in rare or novel scenarios	Continuous safety monitoring and digital twins
Explainability	Opaque "Black-Box" neural architectures	Lack of accountability and failure traceability	Stakeholder-tailored XAI and contrastive logic
Ethics	Dataset bias and lack of moral transparency	Inconsistent safety performance across populations	Unified ethical and governance frameworks

CONCLUSION

This narrative review examined the state of AI safety in autonomous driving. Current studies demonstrate that safety challenges are interconnected system-level issues. Limitations in perception accuracy continue to constrain environmental understanding, while the black-box nature of deep learning models complicates failure diagnosis. Achieving safe autonomous driving cannot rely solely on improving algorithmic performance; robust operation must be complemented by explainable decision-making mechanisms that enable transparency and accountability.

The reviewed literature also highlights that achieving safe autonomous driving cannot rely solely on improving algorithmic performance. Robust operation under uncertainty must be complemented by explainable decision-making mechanisms that enable transparency, traceability, and accountability (11, 12). Safety assurance approaches that integrate robustness analysis with explainability provide more effective support for verification, validation, and certification processes (1, 11, 12). Furthermore, ethical and governance considerations, including fairness, accountability, and cybersecurity, are inseparable from technical safety, as emphasized by recent work on trustworthy AI frameworks for transportation systems (2).

Looking forward, future research should prioritize the development of learning methods that maintain stable and predictable behavior under diverse and evolving real-world conditions (15). Improving dataset diversity and edge-case representation remains essential for strengthening perception reliability (4, 15). Continued advancement in explainable artificial intelligence is required to support regulatory evaluation, post-deployment monitoring, and continuous safety assurance (11, 12). In parallel, clearer and more adaptive regulatory and governance frameworks are needed to guide certification, oversight, and responsible deployment of autonomous vehicle technologies (2). Together, these efforts are critical to enabling autonomous vehicles that are not only intelligent, but also safe, transparent, and aligned with societal values.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the guidance of Hamid Foroughi, a mentor from Johns Hopkins University, in the development of this manuscript.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Wäschle M, Thaler F, Berres A, Pölzlbauer F & Albers A. A review on AI Safety in highly automated driving. *Frontiers in artificial intelligence*. 2022; 5: 952773. https://doi.org/10.3389/frai.2022.952773
2. Acharya DB, Kuppan K & Divya B. Agentic ai: Autonomous intelligence for complex goals-a comprehensive survey. IEEe Access. 2025. https://doi.org/10.1109/ACCESS.2025.3532853
3. Koopman P & Wagner M. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*. 2016; 4 (1): 15-24. https://doi.org/10.4271/2016-01-0128
4. Zhang P, Li X, Lin X & He L. A new literature review of 3D object detection on autonomous driving. *Journal of Artificial Intelligence Research.* 2025; 82: 973-1015. https://doi.org/10.1613/jair.1.15961
5. Goodfellow IJ, Shlens J & Szegedy C. Explaining and harnessing adversarial examples. 2015. arXiv preprint arXiv:1412.6572.
6. Eykholt K, Evtimov I, Fernandes E, Li B, *et al*. Robust physical-world attacks on deep learning visual classification. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018; pp.1625-1634. https://doi.org/10.1109/CVPR.2018.00175
7. Madry A, Makelov A, Schmidt L, Tsipras D & Vladu A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*. 2018.
8. Katz G, Barrett C, Dill DL, Julian K & Kochenderfer MJ. Reluplex: An efficient SMT solver for verifying deep neural networks. In International Conference on Computer Aided Verification. 2017; pp.97-117. Springer. https://doi.org/10.1007/978-3-319-63387-9_5
9. Huang X, Kwiatkowska M, Wang S & Wu M. Safety verification of deep neural networks. In International Conference on Computer Aided Verification. 2017; pp.3-29. Springer. https://doi.org/10.1007/978-3-319-63387-9_1
10. Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, *et al*. AI2: Safety and robustness certification of neural networks with abstract interpretation. In 2018 IEEE Symposium on Security and Privacy (SP). 2018; pp. 3-18. IEEE. https://doi.org/10.1109/SP.2018.00058

11. Atakishiyev S, Salameh M, Yao H & Goebel R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access*. 2024. https://doi.org/10.1109/ACCESS.2024.3431437
12. Kuznietsov A, Gyevnar B, Wang C, Peters S & Albrecht SV. Explainable AI for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*. 2024. https://doi.org/10.1109/TITS.2024.3474469
13. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*. 2019; 267: 1-38. https://doi.org/10.1016/j.artint.2018.07.007
14. International Organization for Standardization. (2022). Road vehicles - Safety of the intended functionality (ISO Standard No. 21448:2022). https://www.iso.org/standard/77490.html
15. Chen S, Liao Y, Wang F, Wang G, *et al*. Toward the robustness of autonomous vehicles in the AI era. *The Innovation*. 2025; 6 (3). https://doi.org/10.1016/j.xinn.2024.100780
16. Kaur K, Singh S & Kumar N. Digital twins for autonomous vehicle safety: A survey on architectures, enablers, and challenges. IEEE Communications Surveys & Tutorials. 2023.
17. Zhou J, Wang H & Zhang L. Simulation-based testing for autonomous driving systems: A review of tools and methodologies. *Journal of Safety Research*. 2024.
18. Parasuraman R & Riley V. Humans and automation: Use, misuse, disuse, abuse. Human Factors. 1997; 39 (2): 230-253. https://doi.org/10.1518/001872097778543886
19. Endsley MR. From automation surprise to immunotherapy: Addressing the human factor in autonomous systems. *Theoretical Issues in Ergonomics Science*. 2017; 18 (4): 345-361.
20. Ribeiro MT, Singh S & Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp.1135-1144. https://doi.org/10.1145/2939672.2939778, https://doi.org/10.18653/v1/N16-3020
21. Lundberg SM & Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS). 2017; 30.
22. Mohseni S, Zarei N & Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS). 2021; 11 (3-4): 1-45. https://doi.org/10.1145/3387166
23. Adadi A & Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018; 6: 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052
24. Li L, Huang WL, Liu Y, Zheng NN & Wang FY. Parallel testing of vehicle intelligence via virtual-real interaction. Science Robotics. 2018; 3 (15): eaar6934.
25. Varshney KR & Alemzadeh H. On the safety of machine learning: Specific challenges of deep learning in safety-critical systems. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017.