RTET- 2K26", National Conference (June-2026)
Organized by Department of General Sciences and Engineering, AITRC Vita

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Are Computerized Self-Driving Cars Safe? A Critical Review of AI-Based Autonomous Vehicle Safety, Reliability, Risks, and Future Challenges

Prof. S. M. Chougule[1], Mr. Om Mane[2], Mr. Balaji Mohite[3], Ms. Harshada Mole[4], Ms. Anuja Mithari[5], Mr. Dada Metkari[6], Ms. Tejasvi Mane[7], Mr. Om Pawar[8], Ms. Megha More[9]

[1]Assistant Professor, Department of General sciences and Engineering, AITRC, Vita.
2-9Students, Department of General sciences and Engineering, AITRC, Vita.

Abstract- Autonomous vehicles (AVs) powered by artificial intelligence promise a fundamental transformation of road transportation, with potential benefits ranging from near-elimination of human-error-induced crashes to enhanced mobility for populations unable to drive. Yet despite substantial engineering progress and billions of kilometres of cumulative testing, the question of whether computerized self-driving cars are genuinely safe remains technically nuanced and far from resolved. This review critically examines the current state of AI-based autonomous vehicle safety by synthesizing findings from three recent peer-reviewed studies: (1) Grewal et al. [2024/2025], who evaluate Bayesian uncertainty quantification (UQ) methods for predicting safety-critical misbehaviours in simulation-based ADS testing; (2) Nouri et al. [2024], who propose an LLM-based pipeline for automating Hazard Analysis and Risk Assessment (HARA) under ISO 26262 and SOTIF; and (3) Ullrich et al. [2024], who survey the intersection of AI safety research, standardization, and regulation for automated vehicles.

Keywords: Autonomous Vehicles; AI Safety; Deep Neural Networks; Uncertainty Quantification; HARA; SOTIF; ISO 26262; Explainable AI; Cybersecurity; Regulatory Frameworks; Edge Cases; Sensor Fusion.

I.INTRODUCTION

The prospect of vehicles navigating public roads without human intervention has captivated engineers, policymakers, and the public for decades. What was once science fiction has become a multi-billion-dollar enterprise, with Waymo, Cruise, Tesla, Mobileye, and dozens of others advancing prototype systems toward commercial deployment. Yet the central safety question—whether AI-controlled vehicles are safer than human drivers—resists a simple answer, and the confidence with which some industry actors have proclaimed imminent full automation has repeatedly outpaced demonstrable technical reality.

Human drivers cause approximately 1.35 million fatalities worldwide per year, with driver error implicated in roughly 94% of serious US crashes [NHTSA, 2018]. This statistic is frequently cited to argue that almost any autonomous system would represent an improvement. The argument has intuitive appeal but conceals important complications. Human error is distributed across an enormous range of scenarios, severities, and contributing factors. Replacing the human driver requires not merely that the

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

AI system perform better on average, but that it degrade gracefully across the long tail of rare, high-stakes situations that collectively account for a disproportionate share of harm. An autonomous vehicle that excels on well-mapped highways but fails catastrophically at unmarked intersections in heavy rain cannot be considered safe regardless of aggregate statistics.

The technical challenges of AV safety involve multiple interacting subsystems: sensor suites that must perceive the environment under adversarial conditions, machine learning models whose behavior cannot be formally verified, planning algorithms that must account for the intentions of unpredictable road users, and fail-safe mechanisms that must activate reliably without false positives. Layered on top of these engineering challenges are regulatory frameworks that are still evolving, ethical dilemmas about how AI should prioritize competing harms, and cybersecurity threats that could be weaponized against vehicles at scale.

This review critically synthesizes the current state of knowledge across these dimensions. Three recent peer-reviewed contributions form its primary source material. Grewal et al. [2024/2025] evaluate Bayesian UQ methods for predicting safety-critical misbehaviours, finding that Deep Ensembles achieve strong failure prediction rates in simulation with computational overhead compatible with real-time deployment. Nouri et al. [2024] propose and evaluate an LLM-based pipeline for HARA that expert safety engineers found comparable to human-produced outputs as a preliminary step but insufficient as a standalone solution. Ullrich et al. [2024] survey the AI safety assurance landscape comprehensively, arguing that the field must shift toward data-driven assurance methodologies to match the characteristics of the AI systems being deployed.

These papers collectively span the technical (runtime monitoring), process (safety engineering), and governance (standards and regulation) dimensions of AV safety. Their synthesis reveals both important advances and significant residual gaps that define the frontier of the field.

II. BACKGROUND OF AUTONOMOUS VEHICLES

The conceptual lineage of self-driving vehicles stretches to the 1939 Futurama exhibit at the New York World's Fair, where General Motors presented visions of electronically guided highway cars. Practical research accelerated through the 1980s and 1990s with Ernst Dickmanns' VaMoRs project in Germany and Carnegie Mellon University's NAVLAB series in the United States. The pivotal modern phase began with the DARPA Grand Challenge (2004–2005) and Urban Challenge (2007), which demonstrated that autonomous navigation of complex real-world environments was achievable and catalyzed the commercial programs that followed.

Google's self-driving project, which became Waymo in 2016, began accumulating real-world miles in 2009 and today represents one of the most mature fully driverless deployments, with commercial robotaxi service in Phoenix, San Francisco, and other cities. Tesla's Autopilot and Full Self-Driving (FSD) suite have brought partial automation to millions of consumer vehicles, though the system remains SAE Level 2 despite its commercial nomenclature, and has been the subject of multiple fatal crash investigations by the NHTSA. Cruise received regulatory approval for driverless operations in San Francisco before suspending operations in late 2023 following a serious pedestrian injury incident, illustrating the fragility of public trust in this technology.

The economic motivation is substantial: McKinsey has projected annual global AV technology revenue of $300–400 billion by 2035. Yet the timeline compression that characterized early projections has repeatedly failed to materialize—Waymo's 2018 claims of imminent broad commercial deployment gave way to a decade-long period of cautious, geofenced expansion. Understanding why the gap

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

between demonstration capability and robust real-world reliability has proven far wider than initial enthusiasm suggested is a central analytical task of this paper.

III. LEVELS OF VEHICLE AUTOMATION

Figure 1: SAE Levels of Vehicle Automation (0–5)

The SAE International J3016 standard defines six levels of driving automation from Level 0 (No Automation) to Level 5 (Full Automation). This taxonomy has become the dominant framework for describing AV capability and is referenced throughout regulatory and technical literature, including all three primary papers reviewed here.

Level 0 encompasses conventional vehicles where the driver performs all tasks. Level 1 provides isolated assistance acting on a single axis (e.g., adaptive cruise control). Level 2 enables simultaneous automation of steering and acceleration but requires continuous driver monitoring—Tesla Autopilot is the most widely deployed example. Level 3 allows the driver to disengage attention during the automated mode but requires readiness to respond to takeover requests; Honda's Traffic Jam Pilot is a rare commercial example approved in Japan and Germany. Level 4 permits full driving within a defined Operational Design Domain (ODD) without driver intervention, as demonstrated by Waymo One. Level 5 envisions handling all conditions a human driver could navigate—a capability that remains, as of 2025, beyond demonstrated reach.

The ODD concept is critically important for safety argumentation. A Level 4 system may encounter Level 0 conditions immediately upon leaving its ODD boundary, and the transition management—what happens when the vehicle is about to exit its ODD—is one of the most difficult engineering and regulatory problems in the field. ISO 21448 (SOTIF) specifically addresses this challenge, and several failure modes analyzed in Grewal et al. [2024/2025] arise precisely at ODD boundaries.

IV. AI TECHNOLOGIES USED IN SELF-DRIVING CARS

Modern AV perception relies on heterogeneous sensor arrays typically including cameras (texture, colour, high-resolution object recognition), LiDAR (precise 3D depth mapping), radar (velocity estimation and adverse-weather penetration), ultrasonic sensors (close-range detection), and high-definition maps updated in real time. The fusion of these modalities is managed by a perception stack whose reliability depends on both individual sensor performance and the robustness of the fusion algorithm.

Camera-centric approaches, championed most visibly by Tesla, argue that human roads are designed for vision-based agents and that cameras provide sufficient information density at low cost. LiDAR-centric approaches, adopted by Waymo and most robotaxi operators, contend that geometric 3D data

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

is essential for safety-critical decisions. This architectural divergence represents an unresolved engineering debate with significant safety implications: camera-only systems are more vulnerable to certain lighting and weather conditions, while LiDAR adds cost and has its own failure modes in heavy precipitation.

The perception and control pipelines of modern AVs are dominated by deep neural networks (DNNs)—most commonly convolutional neural networks (CNNs) for image-based tasks, recurrent architectures and transformers for temporal sequence modeling, and increasingly large language models for scene understanding and trajectory prediction. The NVIDIA DAVE-2 architecture, the object of study by Grewal et al. [2024/2025], exemplifies the end-to-end behavioral cloning paradigm: a DNN trained to predict steering commands directly from camera images, learning an implicit representation of lane-following behavior. End-to-end models demonstrate compelling performance on in-distribution scenarios but face well-documented challenges with out-of-distribution (OOD) inputs—a problem central to the failure analysis in Grewal et al.

The modular ADS architecture decomposes the driving pipeline into perception, prediction, planning, and control modules, each potentially verifiable independently, but introduces its own challenges in interface specification, error propagation, and system integration. As Ullrich et al. [2024] note, both architectural paradigms coexist without a clear performance or safety advantage emerging for either.

An emerging application of AI to AV safety is the use of large language models in the safety engineering process itself. Nouri et al. [2024] demonstrate a pipeline in which GPT-4 is used to automate HARA—generating hazardous event identifications, severity assessments, and safety goals from item definitions. This represents a qualitatively different use of AI: not as the vehicle control system, but as an assistant to engineers responsible for safety argumentation. The implications—efficiency gains, hallucination risks, expertise dependencies—are explored in the comparative analysis of Section 14.

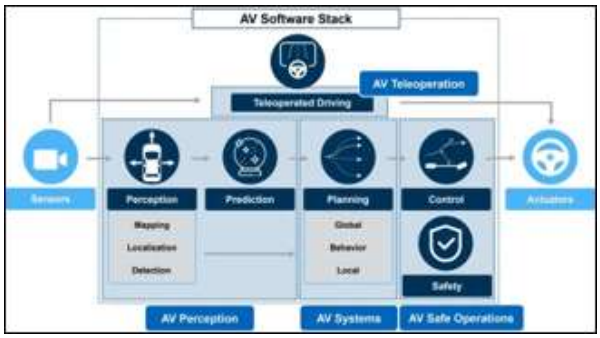

Figure 2: Architecture of an Autonomous Vehicle Perception-Planning-Control

V. SAFETY BENEFITS OF AUTONOMOUS VEHICLES

The primary safety argument for AVs rests on the elimination of human cognitive limitations: fatigue, distraction, intoxication, slow reaction times, and poor hazard perception. A Level 4–5 AV does not fall asleep, does not check a phone, does not experience road rage, and reacts to hazards within milliseconds rather than the human perception-reaction average of approximately 1.5 seconds. On instrumented test tracks under controlled conditions, AV systems consistently outperform average human drivers on reaction time, following distance maintenance, and lane discipline.

Waymo's published safety data from Phoenix and San Francisco deployments is instructive. A 2023 comparison of Waymo's commercial service against human taxi drivers in the same geographic area reported approximately 85% fewer police-reported crashes per vehicle mile traveled. However, such

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

statistics must be interpreted cautiously: deployment environments are carefully selected for AV suitability, the volume of miles may be insufficient for statistically reliable rare-event comparisons, and survivorship bias may affect which scenarios are included in the analysis.

There are also systemic safety benefits beyond individual vehicle performance. AV platoons could dramatically reduce highway fatalities through coordinated sub-second braking. V2X (vehicle-to-everything) communication could allow AVs to collectively detect hazards before any individual vehicle's sensors register them. At population scale, the removal of impaired and distracted drivers—who are disproportionately responsible for serious crashes—would reduce harm even if AI substitutes were merely average rather than superior.

VI. MAJOR SAFETY RISKS AND FAILURES

The safety risks of autonomous vehicles are multidimensional, spanning individual sensor failures, algorithmic edge cases, systemic software bugs, cybersecurity vulnerabilities, and regulatory gaps. A taxonomy of AV failure modes is a prerequisite for structured safety analysis.

High-profile incidents provide concrete anchors for abstract risk categories. The 2016 fatal Tesla Autopilot crash in Florida involved a camera-based system failing to distinguish a white trailer against a bright sky—a classic OOD perception failure that LiDAR might have detected. The 2018 Uber ATG pedestrian fatality in Tempe, Arizona resulted from sensor misclassification, an overly aggressive object filtering algorithm, a miscalibrated decision tree, and an inattentive safety driver—illustrating how multiple independent failures at different abstraction layers compound into catastrophe. The 2023 Cruise pedestrian dragging incident in San Francisco exposed failures in post-collision decision-making algorithms and subsequent incident reporting practices.

These incidents share a common thread: AV systems encountered conditions rare in their training data, at the ODD boundary, or ambiguous enough that the system's confidence was not appropriately calibrated. The uncertainty quantification work of Grewal et al. [2024/2025] directly targets this problem. Their framework is designed to detect such conditions before they produce failures by monitoring DNN confidence during operation and triggering warnings when uncertainty spikes. That Deep Ensembles achieved F3 scores of 90–94% in predicting failures several seconds in advance represents meaningful progress, but the simulation context—Udacity, a relatively simple lane-keeping environment—limits direct extrapolation to real-world deployment.

VII. SENSOR AND PERCEPTION CHALLENGES

Sensor reliability under real-world conditions constitutes one of the most technically demanding aspects of AV safety engineering. Camera systems are susceptible to glare, lens contamination, snow accumulation, and low-light conditions. LiDAR systems can be confused by retroreflective surfaces, heavy rain or snow that scatters beam returns, and shadows cast by other vehicles. Radar is relatively weather-robust but has lower resolution and can struggle with static object discrimination. No single sensor modality is sufficient; the safety argument for perception depends entirely on multi-modal fusion and the assumption that failures in different sensors are uncorrelated.

This independence assumption breaks down in precisely the most dangerous conditions. A dense snowstorm simultaneously impairs cameras, LiDAR, and radar while also degrading GPS reliability. The epistemic uncertainty formalized by Grewal et al. [2024/2025]—arising from the system's lack of knowledge about conditions outside its training distribution—manifests in exactly these multi-sensor degradation scenarios. Their OODextreme benchmark, which introduces simultaneous night and snow

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

conditions, is instructive: Deep Ensembles maintained high failure prediction accuracy even under severe distributional shift, suggesting that uncertainty monitoring provides a safety layer precisely when sensors are most challenged.

ISO 21448 (SOTIF), discussed extensively by Ullrich et al. [2024] and forming the backbone of Nouri et al.'s [2024] HARA framework, provides a structured methodology for identifying and mitigating perception-related safety risks—specifically hazards arising not from component malfunction but from functional insufficiency in conditions the system was not designed for. The standard distinguishes between known unsafe scenarios (for which mitigation is straightforward), unknown unsafe scenarios (the genuinely difficult case), and the triggering conditions that produce unsafe behavior. Closing the gap between estimated and actual unknown unsafe scenarios is the central challenge of AV safety validation.

VIII. EDGE CASES AND REAL-WORLD DRIVING COMPLEXITY

The concept of an edge case in autonomous driving requires precise definition. What matters for safety is not merely low prior probability but the interaction between scenario probability, failure severity, and the system's safety margin when encountering that scenario. A zebra crossing in light rain is statistically uncommon compared to dry-weather highway driving, but common enough in real operations and consequential enough that reliable handling is non-negotiable.

The challenge is not simply enumerating edge cases—any practitioner can generate lists of difficult scenarios—but developing a systematic methodology for achieving coverage. ISO 21448 proposes scenario-based testing with systematic catalogues, and the PEGASUS research project has developed a six-layer model for urban traffic scenario description. Nouri et al. [2024] confront this problem directly in their HARA pipeline: experts reviewing LLM-generated HARAs cited incompleteness of scenario coverage as a primary limitation, with one reviewer noting that identifying all relevant scenarios 'is more like a craft and no systematic way exists till now.'

This admission from an expert automotive safety engineer is significant. It suggests that the completeness of safety argumentation for AV systems remains at least partially a matter of judgment rather than provable specification—which has profound implications for regulatory approval. A vehicle cannot be certified as safe if the argument for the completeness of hazard coverage is, by expert admission, craft rather than science. The LLM pipeline proposed by Nouri et al. offers efficiency gains in hazard generation but does not resolve the completeness problem; it may, however, make the search space more systematically explorable by enabling rapid generation and review of candidate hazardous events at scale.

REFERENCES

1. R. Grewal, P. Tonella, and A. Stocco, "Predicting Safety Misbehaviours in Autonomous Driving Systems using Uncertainty Quantification," arXiv:2404.18573v2 [cs.LG], Feb. 2025.
2. A. Nouri, B. Cabrero-Daniel, F. Torner, H. Sivencrona, and C. Berger, "Engineering Safety Requirements for Autonomous Driving with Large Language Models," Proc. 32nd IEEE International Requirements Engineering Conference (RE 2024), Iceland, arXiv:2403.16289v1, 2024.
3. L. Ullrich, M. Buchholz, K. Dietmayer, and K. Graichen, "AI Safety Assurance for Automated Vehicles: A Survey on Research, Standardization, Regulation," IEEE Transactions on Intelligent Vehicles, arXiv:2504.18328v1, Nov. 2024, doi: 10.1109/TIV.2024.3496797.
4. SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," SAE Standard J3016_202104, 2021.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

5. ISO, "Road Vehicles — Functional Safety," ISO 26262:2018, International Organization for Standardization, 2018.
6. ISO, "Road Vehicles — Safety of the Intended Functionality," ISO 21448:2022, International Organization for Standardization, 2022.
7. ISO/SAE, "Road Vehicles — Cybersecurity Engineering," ISO/SAE 21434:2021, 2021.
8. 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.
9. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Proc. 33rd ICML, vol. 48, pp. 1050–1059, 2016.
10. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," Advances in NeurIPS, vol. 30, 2017.
11. A. Stocco, M. Weiss, M. Calzana, and P. Tonella, "Misbehaviour Prediction for Autonomous Driving Systems," Proc. 42nd ICSE, ACM, 2020.
12. A. Stocco, P. J. Nunes, M. d'Amorim, and P. Tonella, "ThirdEye: Attention Maps for Safe Autonomous Driving Systems," Proc. 37th IEEE/ACM ASE, 2022.
13. D. Amodei, C. Olah, J. Steinhardt et al., "Concrete Problems in AI Safety," arXiv:1606.06565, 2016.
14. National Highway Traffic Safety Administration (NHTSA), "Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey," DOT HS 812 506, 2018.
15. European Parliament, "Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act)," adopted March 13, 2024.
16. National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
17. P. Koopman, U. Ferrell, F. Fratrik et al., "A Safety Standard Approach for Fully Autonomous Vehicles," Proc. 38th SAFECOMP, Springer, pp. 326–332, 2019.
18. E. Hullermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," Machine Learning, vol. 110, pp. 457–506, 2021.
19. M. Weiss and P. Tonella, "Fail-safe Execution of Deep Learning Based Systems through Uncertainty Monitoring," Proc. 14th IEEE ICST, 2021.
20. N. Humbatova, G. Jahangirova, G. Bavota et al., "Taxonomy of Real Faults in Deep Learning Systems," Proc. 42nd ICSE, ACM, 2020.
21. W. He and Z. Jiang, "A Survey on Uncertainty Quantification Methods for Deep Neural Networks," 2023.
22. T. Brown, B. Mann, N. Ryder et al., "Language Models are Few-Shot Learners," Advances in NeurIPS, vol. 33, pp. 1877–1901, 2020.
23. ANSI/UL, "Standard for Evaluation of Autonomous Products," ANSI/UL 4600, 2020.
24. S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A Survey of Deep Learning Techniques for Autonomous Driving," Journal of Field Robotics, vol. 37, no. 3, pp. 362–386, 2020.
25. A. V. S. Neto, J. B. Camargo, J. R. Almeida et al., "Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review," IEEE Access, vol. 10, pp. 130733–130770, 2022.
26. M. Scholtes et al., "6-Layer Model for a Structured Description and Categorization of Urban Traffic and Environment," IEEE Access, vol. 9, pp. 59131–59147, 2020.
27. M. Bojarski et al., "End to End Learning for Self-Driving Cars," arXiv:1604.07316, 2016.
28. ISO, "Road Traffic Safety — Guidance on Ethical Considerations Relating to Safety for Autonomous Vehicles," ISO 39003:2023, 202.