**International Journal of Emerging Trends in Computer Science and Information Technology**
ISSN: 3050-9246 | https://doi.org/10.56472/WCAI25-120
Eureka Vision Publication | WCAI25-Conference Proceeding

*Original Article*

# Design and Evaluation of AI Safety Mechanisms in ADAS and Autonomous Vehicle Architectures

Gaurav Pokharkar
Product Technial Leader, Valeo, USA.

*Abstract - The progressive shift from human-driven to automated driver assistance and to fully autonomous vehicles has placed Artificial Intelligence (AI) at the center of automotive innovation. Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicle (AV) platforms increasingly depend on AI for perception, decision making, and controls leading to enhanced vehicle safety and vehicle operational efficiency. However, the non-deterministic nature of AI, coupled with its dependence on training data and susceptibility to out-of-distribution (OOD) inputs, introduces novel safety hazards not encountered in traditional deterministic control systems [1] [2]. This paper aims to investigate the design and evaluation of safety mechanisms capable of detecting, mitigating, and recovering from unexpected AI behaviors in OOD scenarios for the AI system. The study considers layered safety architectures, continuous monitoring strategies, dataset lifecycle management, simulation-based validation, and performance metric analysis as part of an integrated safety framework. Key findings include that OOD detection techniques, such as Mahalanobis distance-based scoring, can significantly reduce misclassifications risk, although sometimes at the expense of operational coverage [3]. The integration of safety monitors into perception pipelines has been shown to improve system trustworthiness by identifying failure patterns before they escalate into hazardous decisions [4]. Furthermore, empirical studies reveal demographic performance disparities in pedestrian detection models, particularly under low-light or low-contrast scenarios, which highlights the importance of bias aware dataset curation [5] [6]. We conclude that achieving resilient autonomy demands a multi-layered safety approach: combining proactive monitoring, fallback control logic, diverse and bias mitigated datasets, rigorous simulation-based testing, and alignment with evolving regulatory standards. These measures form the basis for developing trustworthy AI systems capable of operating safely in unpredictable scenarios encountered during real-world driving conditions.*

*Keywords - Index Terms, ADAS, Autonomous vehicles, AI safety, OOD detection, Bias mitigation, Simulation validation, Functional safety.*

## 1. Introduction

Over the past decade, the automotive industry has undergone a rapid transformation, mostly driven by advancements in Advanced Driver Assistance Systems (ADAS) eventually with a shift toward fully autonomous vehicle (AV) capabilities. ADAS functions such as adaptive cruise control, lane keeping assistance, automated lane change, and collision mitigation are now widely available in production vehicles with various OEMs. This represents the first wave of automation in consumer transport [7] [8]. The goal of these systems is to enhance driver safety, reduce collisions, and improve traffic efficiency, while maintaining a human operator in control and final decision maker in the vehicle. As we move from ADAS to high-level automation as shown in Figure 1 it leads to a shift in system design strategy utilizing Artificial Intelligence (AI) at the core of perception, decisionmaking, and control. AI driven perception stacks, employing deep neural networks, have shown remarkable capability in recognizing objects, road signs and predicting the behavior of surrounding traffic [9]. However, this reliance on AI introduces new safety considerations absent in traditional deterministic vehicle control systems.

The benefits of AI in AVs are significant it enables human like adaptive behavior in complex environments, it helps improve detection under variable conditions, and can process multi-modal sensor data more effectively than deterministic algorithms [10]. However, these strengths are counterbalanced since neural networks are inherently data dependent, their ability to generalize is limited by the diversity and representativeness of training datasets [11]. Failures can arise in edgecase scenarios, such as unusual weather, low light conditions, construction zones, or rare pedestrian behaviors, where the AI encounters out-of-distribution (OOD) inputs [3]. Also, AI models often operate as “black boxes,” with decisionmaking processes that are difficult to verify or explain which complicates both development and post-incident analysis [12]. These challenges are compounded by the unpredictability of real-world driving. There are infinite range of scenarios encountered while driving due to the combination of vehicle types, weather, zones, road types etc.

Although conventional safety engineering in automotive systems has long addressed deterministic hardware and software failures, the stochastic and context-dependent nature of AI failures demands new approaches. Existing standards such as ISO 26262 and ISO/PAS 21448 (SOTIF) address functional safety and safety of the intended functionality, yet their adaptation to AI-intensive architectures remains an evolving field [1] [2]. The purpose of this paper is to analyze how AI safety mechanisms can be designed, evaluated, and validated to handle failures in real-world driving contexts. We examine layered architectures, continuous monitoring, fallback control strategies, dataset lifecycle management, simulation-based validation, and performance metrics, with the aim to provide a comprehensive framework for developing trustworthy, regulation-compliant AI systems for ADAS and AVs.

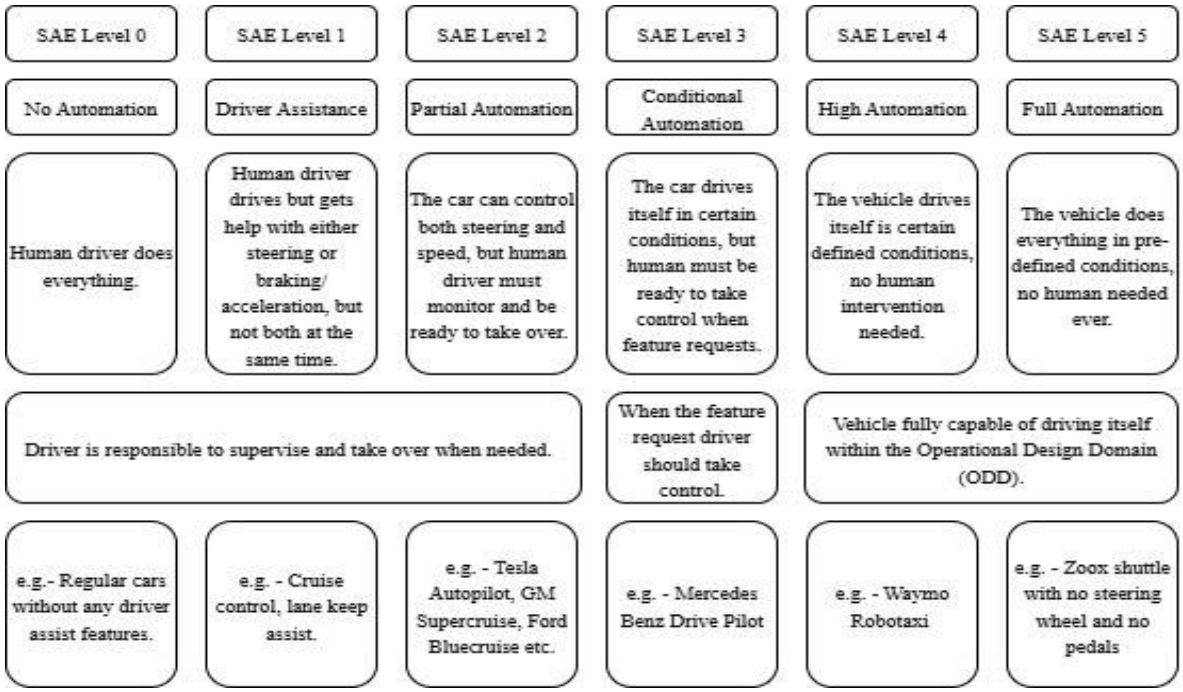


**Figure 1. SAE Automation Levels [26]**

## 2. Safety Hazards Unique To Ai-Driven Adas/Avs

### *2.1. Data Bias and Demographic Disparities*

A well known risk in AI based perception is due to dataset bias. Pedestrian detection models, trained predominantly on improperly balanced datasets, may perform poorly on specific demographic groups. Some studies have shown that detection accuracy is significantly lower for children and individuals with darker skin tones, particularly under challenging illumination conditions such as low light or fog [6] [5]. These gaps aren't just a research problem, the bias in safety-critical models can directly put certain road users, especially vulnerable ones, at greater risk. The Predictive Inequity in Object Detection study [5] demonstrated that, in a widely used pedestrian dataset, the detection confidence was consistently lower for pedestrian with darker skin tone. Environmental factors such as low light and low contract conditions compounded this effect leading to sharp increase in the false negatives and also recognition delays lengthen [13]. Such findings underscore the need for deliberate dataset curation, augmentation, and bias testing as part of the safety assurance process.

### *2.2. Out-of-Distribution (OOD) Inputs and Model Uncertainty*

Deep neural networks operate reliably within the statistical limits of their training data. When confronted with out-ofdistribution (OOD) input such as construction zone, an irregular traffic pattern, or environmental conditions that are not encountered during training as shown in Figure 2, the behavior of the model can degrade unpredictably [14]. In autonomous vehicles, this can manifest itself as misclassifications, missed detections, or unsafe trajectory proposals.Techniques such as Mahalanobis distance-based scoring [3] and uncertainty estimation through ensemble methods [15] have been shown to improve OOD detection, allowing the system to trigger safe fallback actions. However, these methods involve trade-offs: lowering the detection threshold to catch more OOD cases can unnecessarily restrict operational coverage, impacting usability.

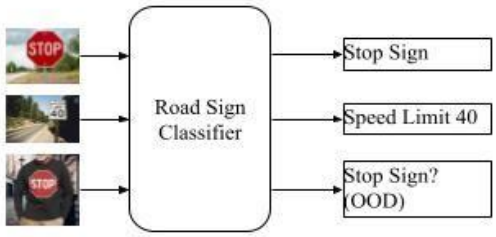


**Figure 2. Out of Distribution Detection**

### *2.3. lack-Box Decision Processes*

Most high-performing perception and planning models in AVs are based on deep learning architectures, which inherently lack transparency. Their decision boundaries and feature attributions are difficult to interpret without specialized tools [12]. This opacity complicates verification, validation, and root-cause analysis after incidents. In safety-critical domains, the inability to provide an audit trail for decisions is a barrier to both regulatory compliance and public trust. Explainable AI (XAI) methods such as saliency mapping, feature attribution, and activation analysis are being explored to improve interpret ability [16]. Although promising, these methods are not yet standardized for automotive safety certification.

### *2.4. Sensor Fusion Fragility*

Multi-modal sensor fusion (e.g., combining camera, radar, and LiDAR inputs) as shown in as shown in Figure 3 is intended to improve perception robustness by leveraging complementary sensing modalities. However, faults in one modality can propagate through the fusion pipeline, contaminating downstream perception and planning layers. For example, misaligned calibration in a camera-LiDAR pair can cause object position errors, leading to unsafe path planning [17].Failure modes in sensor fusion can arise from hardware degradation, environmental interference (e.g., radar multi-path in urban canyons), or software errors in the fusion algorithm itself. Because fusion is deeply integrated into decision making, even subtle sensor-specific anomalies can have amplified safety consequences.

### *2.5. Real-World Failure Example*

A stark illustration of these hazards is found in the 2018 Tempe, Arizona fatality involving an AV test vehicle. Postincident analysis indicated that the object detection system failed to correctly classify a pedestrian crossing outside of a crosswalk until it was too late to initiate an avoidance maneuver [18]. The misclassification by the system, combined with an insufficient fallback response, exemplifies the convergence of several hazards: biased training data (infrequent pedestrianoutside-crosswalk scenarios), OOD conditions (dimly lit environment), black-box decision opacity, and sensor fusion limits. This case emphasizes the need for AI safety mechanisms that can not only detect anomalies but also respond rapidly and deterministically in ways that preserve life.

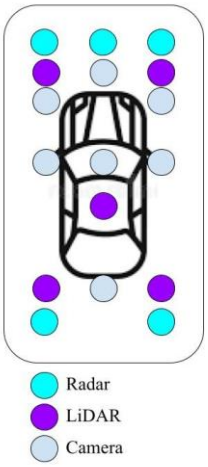


**Figure 3. Multi Modal Sensor Fusion**

## 3. Ai Safety Mechanisms: Design Principles

The design of AI safety mechanisms for ADAS and AVs requires a multi-pronged approach that addresses failure prevention, detection, and recovery. Safety must be embedded into both the architectural foundations and the runtime monitoring of AI components. This section outlines key design principles organized under safety-oriented AI architecture and control and monitoring strategies.

### *3.1. Safety-Oriented AI Architecture*

*3.1.1. Redundant Sensing*

Robust perception in AVs hinges on multi-modal sensing as shown in as shown in Figure 3. Cameras, radar, and lidar each have complementary strengths: cameras provide high-resolution semantic detail, radar maintains reliability in poor weather, and lidar delivers precise 3D spatial mapping [19]. By fusing these modalities, the system can maintain perception fidelity when one sensor is degradedradar, for instance, can detect vehicles through fog where camera performance suffers [20]. To further reduce systemic risk, redundant sensing can include heterogeneous hardware from multiple suppliers, reducing the chance that a shared defect cascades across all modalities.

*3.1.2. Parallel AI Pipelines*

Running independent perception or planning models in parallel allows for cross-validation of outputs. Discrepancies between pipelines can flag possible faults, prompting a safety review or the fallback action [21]. These models may differ in architecture, training data, or feature representation, ensuring that they fail under different conditions. For example, a convolutional neural network (CNN) and a transformer-based detector can complement each other's weaknesses.

*3.1.3. Safety/Shadow Controller*

A safety or a shadow controller operates in parallel to the main autonomy stack, using deterministic, formally verifiable algorithms to monitor AI outputs and override them if hazardous behavior is detected [22]. The shadow controller does not rely on complex learned representations; instead, it enforces physical constraints, braking distances, and collision avoidance heuristics. For example, if the main AI proposes a lane change that would violate a safe gap threshold, the shadow controller can veto the maneuver and maintain the lane position.

### *3.2. Control and Monitoring Strategies*

*3.2.1. Real-Time Monitoring:*

OOD Detection: Detecting when an input lies outside of the learned distribution of the AI model is critical to avoid unsafe extrapolation. Methods like Mahalanobis distance based scoring [3] and softmax entropy maps [23] allow real-time detection of novel inputs. Upon detection, the system can reduce the autonomy level, slow the vehicle, or transfer control to a human operator. Safety Monitors for Perception: Safety monitors evaluate the AI output against known failure signatures. For example, if bounding boxes jitter excessively between frames or a tracked object disappears prematurely, the monitor can treat this as a sign of perception instability [24]. In such cases, fallback actions such as increasing the following distance or halting lane change can be triggered. Logging and Observability: Recording internal AI states, intermediate feature maps, and decision output is essential for post-incident analysis. High-fidelity logs enable developers and regulators to reconstruct failure chains and improve models [25]. In some safety frameworks, vehicle storage is supplemented by encrypted remote uploads for redundancy.

*3.2.2. Active Mitigation and Recovery:*

Graceful Degradation: When partial system impairment occurs, the AV should degrade functionality in a controlled way. This can involve reducing speed (e.g. tunnel speed mode in GPS-denied environments), limiting autonomy to simpler maneuvers, or prompting a human takeover [26] this behavior prevents abrupt loss of control and extends the decision window for safe recovery. Emergency maneuvers: If critical hazards are detected, such as complete perception failure or a high likelihood of collision, the system's fallback logic can execute model predictive maneuvers for rapid deceleration, evasive swerving within safety limits, or controlled pullover [27]. These maneuvers must be pretested under a variety of conditions to ensure stability and prevent secondary hazards. Human-on-Loop supervision: Some AV platforms incorporate remote operators who can intervene in rare but high-risk scenarios [28]. The AI system must be designed to transmit the situational context in real time, enabling the human supervisor to issue corrective commands without delay.

*3.2.3. Health Assessment and Watchdog Functions:*

Confidence estimation: Ensemble models [15] and self modeling architectures [29] can provide a measure of predictive uncertainty. By quantifying confidence in its own output, the system can dynamically adjust the autonomy level with high confidence states allowing normal operation, while low confidence states trigger cautionary behaviors. Watchdog timers and crosschecks: Watchdog timers detect system hangs or excessive processing delays, initiating a safe stop if the autonomy stack

becomes unresponsive. Cross-checking between processors (e.g. dual redundant compute units running safety critical logic) helps detect erratic output caused by software or hardware faults [30].Self-testing: Routine injection of internal test cases, such as simulated obstacle appearances, can validate that perception, planning, and actuation components work correctly during operation. Failures in these self-tests prompt immediate transitions to safe fallback modes [31].These design principles collectively aim to create AV systems that are resilient to the diverse and unpredictable failure modes inherent to AI-driven autonomy. By combining architectural redundancy, continuous monitoring, active mitigation, and robust health checks, developers can significantly reduce the risk of catastrophic failures.

## 4. Data, Development, and Validation for Ai Safety

Ensuring AI safety in ADAS and AV systems depends on a rigorous approach to data management, simulation testing, and the use of well-defined performance indicators. Since AI models are only as robust as the data and validation processes they are built upon, the pipeline must be explicitly designed to capture real-world diversity, systematically test system responses, and quantify safety performance.

### *4.1. Dataset Life-cycle Safety*

*4.1.1. Diversity in Data Collection*

The resilience of AI in AVs depends on exposure to a wide variety of operational conditions. Standard driving datasets often over represent clear weather, daytime and majority of the demographic conditions [26]. Safety-oriented datasets must intentionally capture rare weather patterns (snow, heavy rain, fog), low illumination settings, and unusual traffic behaviors (jaywalking, sudden lane changes), etc. Inclusion of underrepresented demographic groups is critical, as pedestrian detection has been shown to have reduced accuracy for children, elderly individuals, and people with darker skin tones [32]. These disparities worsen under low light or fog conditions, compounding the risk in real deployments [5]. To fill in natural data gaps, synthetic augmentation techniques, such as generative adversarial networks (GANs) and domain randomization, can simulate challenging conditions, including adversarial lighting or occlusions [33]. Synthetic data cannot fully replace real-world captured data, but can extend coverage across rare scenarios, aiding model generalization.

*4.1.2. Life-cycle management and Annotation*

Dataset management driven by safety is not static; it is an evolving process tied to field performance. Continuous logging data from deployed fleets allows developers to identify performance regressions and emerging edge cases [25]. Annotation processes should include labels for safety critical tags that indicate the potential for harm if misclassified, such as a child on the road or an approaching emergency vehicle. Coverage monitoring tools can track the presence or absence of specific conditions in training data. If foggy nighttime intersections are underrepresented, targeted collection campaigns can be deployed to close these gaps before the next model iteration.

*4.1.3. Bias Checks and Fairness Audits*

Data bias must be evaluated at multiple stages of the lifecycle. Cross-domain bias checks ensure that models trained in one geography generalize to others, while cross-demographic audits detect disparities in detection accuracy between age, gender, and ethnicity [34]. Automated fairness metrics can flag detection or classification discrepancies that exceed predefined thresholds [35]. Furthermore, model retraining should integrate mitigation strategies when bias is identified, such as reweighting loss functions for underrepresented classes or using fairness-constrained learning [36]. Safety validation is incomplete without proving that the model performs equitably across diverse users and environments.

### *4.2. Simulation-Based Validation*

*4.2.1. Scenario Generators*

Simulation is essential for testing AI safety mechanisms in ADAS and autonomous vehicles because it enables exploration of dangerous or rare scenarios that cannot be reproduced safely on public roads. Scenario generators can vary conditions such as lighting, weather, traffic, and road layout in the Operational Design Domain (ODD) [37]. For example, a simple case of a pedestrian crossing mid-block can be programmatically altered by adjusting walking speed, clothing color, visibility, or occlusions. This combinatorial approach ensures that corner cases, including those absent from the collected datasets, are examined prior to deployment.

Simulation-in-the-Loop (SiL) extends this capability by offering a scalable, high-fidelity environment tailored for production intent AI software. Traditional road testing is expensive and poorly suited for capturing rare edge cases, whereas SiL provides cost-effective coverage and faster iteration [41]. By simulating critical edge-case behaviors, developers can identify weaknesses long before vehicles face them on real roads.

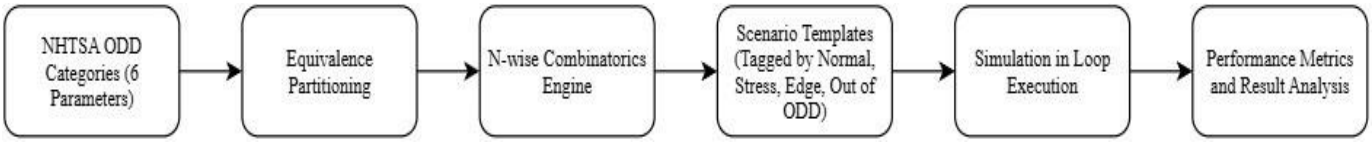


**Figure 4. Scenario Generation Pipeline [41].**

The NHTSA ODD taxonomy [42] is central to this process, separating scenarios by factors such as infrastructure, environmental conditions, and operational constraints. This categorization enables tests across normal, stress, edge, and out-of-ODD conditions as shown in Figure 4, ensuring that AI systems not only handle expected inputs, but also degrade gracefully in abnormal ones. To control the explosion of test combinations, methods such as pairwise testing and equivalence partitioning are used, thereby allowing a broad but efficient scenario coverage [41].SiL also supports agile development cycles. Automated regression testing verifies that new AI releases, including Over-the-Air (OTA) updates, do not compromise safety-critical functions. Version-controlled simulation campaigns ensure traceability [41], an important requirement for standards such as ISO 26262 [1].Ultimately, SiL is more than a validation tool; it provides a structured framework for early bug detection, accelerated development, and regulatory alignment [41]. By embedding safety checks throughout the software lifecycle, simulation ensures that AI-driven autonomy advances without sacrificing reliability or public trust.

*4.2.2. Digital Twins and Hardware-in-the-Loop (HIL)*

Digital twin systems replicate real-world locations, road geometry, and traffic conditions with high fidelity, enabling locationspecific safety validation [38]. When combined with HIL setups, these simulations connect the virtual environment to the physical AV hardware stacksensors, compute units, and actuators. This enables validation of not only AI perception and planning but also hardware timing, sensor synchronization, and actuator latency. Integration testing in a digital twin ensures that perception models can handle sensor dropout, planning modules can respond within latency budgets, and control layers can execute safe maneuvers even under degraded conditions [39].Simulation-based testing also provides a mechanism for regression analysis. When a safety incident occurs in realworld deployment, it can be replicated and varied in simulation to test candidate fixes before fleet wide roll out. This reduces the risk of introducing new vulnerabilities during patching.

*4.2.3. Metrics and Performance Indicators*

Safety performance must be quantified using clear and measurable Key Performance Indicators (KPIs) that align with functional safety targets. Common KPIs include:

- Detection rates for critical classes (e.g., pedestrians, cyclists, and emergency vehicles).
- False positive / negative rates, with an emphasis on minimizing safety-critical misses.
- Perception latency, measured from sensor capture to classified output.
- Reaction time, from hazard detection to initiation of a safe maneuver.
- Safe maneuver success rate, evaluating whether the AV executed an intended safety maneuver without secondary hazards.
- Coverage metrics indicating the percentage of known edge cases that are handled successfully [40].

These KPIs must be tracked not only in pre-deployment testing but also during post-deployment monitoring, ensuring that drift or environmental shifts in the real world do not erode safety performance over time.

## 5. Contemporary Implementations and Case Studies

The implementation of AI safety mechanisms in ADAS and AV systems has evolved through a combination of regulatory requirements, OEM-driven safety strategies, and lessons learned from real-world deployment. Modern approaches reflect a layered defense philosophy that combines redundancy, fallback controllers, continuous fleet monitoring, and compliance with formal testing standards.

### *5.1. Redundancy and Shadow Controllers*

To meet Automotive Safety Integrity Level D (ASIL-D) requirements as defined by ISO 26262, leading OEMs integrate redundant sensing and control architectures. Sensor redundancy often pairs complementary modalities camera, radar, and LiDAR, so that environmental perception remains functional if one sensor fails [1]. In adverse weather, radar can retain the ability to detect objects where cameras struggle, and LiDAR can help resolve range ambiguities [43]. Redundancy extends beyond sensing into computation. Shadow controllers, deterministic backup modules run in parallel with AI-driven decision-making. While the primary AI model handles complex perception and planning, the shadow controller continuously evaluates the state of the vehicle against

deterministic safety rules. For example, if lane markings are lost and no safe alternative is identified within a set time window, the shadow controller can initiate a controlled deceleration [25]. This approach allows for graceful degradation without relying entirely on human intervention in rapidly evolving hazards.

### *5.2. Lane-Keeping Fallback Strategies*

Lane-keeping assist (LKA) systems are among the most widely deployed ADAS features. Contemporary implementations use multi-model pipelines that combine convolutional neural networks for visual lane recognition with geometric models to track lane curvature [44]. When the primary perception system loses confidence, due to occlusions, poor weather, or faded markings, the fallback logic is triggered. This may involve requesting immediate driver takeover through visual, auditory, and haptic alerts. If no response is detected, the system can autonomously slow the vehicle, maintaining lateral stability until a stop is reached [45]. Some OEMs use confidence thresholds derived from real-time softmax entropy or OOD detection to determine when to switch to fallback mode [3].

### *5.3. Fleet Health Monitoring*

Post-deployment fleet monitoring is now a core safety practice. Vehicles are equipped with telematics and over-theair (OTA) update capabilities that allow OEMs to continuously capture edge case data encountered in the field [46]. This data is fed back into the AI model retraining pipelines, ensuring that unusual situations such as novel vehicle types or rare weather are addressed in future updates. Health monitoring systems also track component level performance, including sensor calibration drift, abnormal actuator response times, and software process health through watchdog timers [47]. When anomalies are detected, remote diagnostics can trigger pre-emptive maintenance or software rollback. This proactive approach reduces the likelihood of safety-critical failures between scheduled service intervals.

### *5.4. Real-World Effectiveness Studies*

Empirical evaluations indicate that ADAS technologies can measurably reduce collision rates. A large-scale insurance telematics study found that forward collision warning and lane departure warning systems were associated with an average 15% reduction in relevant crash rates, while blind spot monitoring reduced lane change related crashes by approximately 19% [48]. However, these benefits can be tempered by behavioral adaptation, a phenomenon in which drivers adjust their behavior in response to perceived safety nets. For example, certain urgent alerts have been associated with a 5.7% increase in harsh braking events, possibly reflecting over reliance or startling responses [49]. This underscores the importance of designing safety interventions that support, rather than disrupt, driver situational awareness.

### *5.5. Regulatory Testing and Market Surveillance*

ADAS and AV safety claims are increasingly validated through independent testing bodies and harmonized regulations. The European New Car Assessment Program (Euro NCAP) now includes a dedicated Highway Assist rating, which evaluates the interaction between driver assistance features, safety backstops, and human-machine interfaces [50]. United Nations Economic Commission for Europe (UNECE) regulations such as UN R157 for Automated Lane Keeping Systems (ALKS) define performance limits, fallback requirements, and cybersecurity provisions for Level 3 autonomy [51]. Market surveillance mechanisms allow regulators to assess deployed systems after market, ensuring continued compliance with safety standards. These frameworks not only guide OEM implementation, but also serve as benchmarks for public trust. As higher levels of automation become commercially viable, such regulatory scaffolding will be essential to harmonize safety expectations across jurisdictions.

## 6. Standards, Regulatory Framework and Open Challenges

### *6.1. Functional Safety and SOTIF Foundations*

ISO 26262 establishes the functional safety baseline for road vehicles, focusing on avoiding hazards due to electrical or electronic failures [1]. It defines Automotive Safety Integrity Levels (ASIL) from A (lowest) to D (highest), requiring rigorous verification, redundancy, and fault tolerance in safety critical components. Although essential, ISO 26262 mainly addresses systematic and random hardware/software faults, not the performance limitations of AI-based perception and decision systems. To address this gap, ISO/PAS 21448, also known as Safety of the Intended Functionality (SOTIF), extends safety considerations to scenarios where the system functions as designed but may still cause harm, such as failing to detect a pedestrian under unusual lighting conditions [2]. SOTIF emphasizes the identification of unknown hazards during the concept and validation phases and requires scenario-based testing, both in simulation and on-road.

### *6.2. Evolving International Regulations*

The United Nations Economic Commission for Europe (UNECE) has developed UN Regulation No. 157, which governs Automated Lane Keeping Systems (ALKS) at SAE Level 3 [51]. The regulation specifies operational domain limits, fallback requirements, minimum risk maneuvers, and data storage for crash reconstruction. In the European Union, the type approval

framework integrates these UNECE rules into the market entry requirements. It also mandates market surveillance, ensuring that post-sale vehicles continue to comply with safety standards [52]. This framework is expanding to address higher automation levels, integrating cybersecurity (ISO/SAE 21434) and AI ethics guidelines.

### *6.3. Open Challenges*

- **Scaling to new architectures:** Transformer-based models, which excel in computer vision and multimodal fusion, introduce both opportunities and challenges. Their high capacity improves perception, but can exacerbate overfitting to biased datasets and make real-time explainability harder [53].
- **Anticipating "unknown unknowns":** Even with extensive OOD detection and simulation, AVs can encounter unprecedented hazards. AI safety research must explore meta-learning and continual adaptation strategies that generalize beyond known training distributions [54].
- **Explainability for trust:** Both regulators and the public require interpretable output from AI systems. Current post hoc saliency maps and feature attribution methods often lack actionable precision for safety validation [55]. Transparent decision logs and hybrid rule-based checks could bridge this gap.
- **Fairness under varied conditions:** Demographic disparities in pedestrian detection remain a pressing equity concern. Studies have shown reduced detection rates for darkskinned pedestrians, particularly in low light conditions [32]. This undermines both public trust and regulatory acceptance, motivating fairness-aware training pipelines and performance parity benchmarks in safety validation.

## 7. Discussion and Conclusion

This study analyzed the design, evaluation and validation of AI safety mechanisms for ADAS and AV systems. Several findings emerge:

- OOD detection (e.g., Mahalanobis distance, softmax entropy maps) effectively reduces classification risk in perception modules, but often at the cost of operational coverage. Systems must balance false rejections against safety benefits [54].
- Safety monitors- shadow controllers, fallback logic, and watchdog timers significantly improve system dependability, providing deterministic recovery paths when AI fails [51].
- Bias in pedestrian (e.g. child pedestrian as shown in Figure 5) detection not only undermines ethical equity, but also increases collision risk. Brightness contrast post-processing, along with dataset balancing, has shown promise in improving both fairness and accuracy under adverse conditions [32] [56].

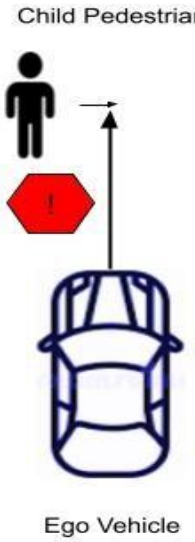


**Figure 5. Bias against child pedestrian**

These results align with previous efforts to improve fairness [32] and confirm the importance of multilayer safety monitoring, as noted in other safety surveys of AV [54]. However, alternative interpretations must be considered. For example, driver adaptation to ADAS features, documented as increased harsh braking in some urgent alert scenarios, suggests that human behavior can offset technical safety gains [49].

### *7.1. Limitations*

Three main constraints shape these findings:

- Simulation vs. real-world gap: Many validation results rely on digital twins that may fail to capture unpredictable driver pedestrian interactions.
- OOD detection trade-offs: While improving safety, high sensitivity can lead to operational conservatism, reducing usability.
- Dataset bias limitations: Bias assessments depend on the availability and diversity of labeled datasets, which may not represent rare demographics or environmental conditions.

### *7.2. Future Research Directions*

To advance trustworthy autonomy, research should focus on the following:

- Adaptive fairness models that dynamically re-weight perception outputs under uncertain conditions.
- Improved OOD detection using multimodal uncertainty fusion.
- Transparency tools designed for regulatory audit and public communication.
- Integration of real-time driver behavior monitoring to dynamically adjust intervention thresholds.
- Regulatory proof-of-concept pilots that test AI safety mechanisms against standardized, yet evolving, scenario suites.

### *7.3. Acknowledgments*

The authors acknowledge the contributions of industry practitioners, regulatory bodies, and open access researchers whose datasets and evaluation frameworks have informed this work. The authors acknowledge the use of OpenAI (GPT5 model, 2025) and Grammarly to assist in improving the clarity, organization, and structure of this manuscript. The AI tools were used strictly for language refinement, grammar correction, and formatting enhancement.

## References

[1] ISO 26262, *Road vehicles Functional safety*, 2nd ed., International Organization for Standardization, 2018.
[2] ISO/PAS 21448, *Road vehicles Safety of the intended functionality (SOTIF)*, International Organization for Standardization, 2019.
[3] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," *arXiv preprint*, arXiv:1807.03888, 2018.
[4] M. Mohseni, M. Pitropov, L. P. Clarke, and T. H. Fung, "Monitoring and Improving Neural Network Safety Using Outlier Exposure," *arXiv preprint*, arXiv:2412.06869, 2024.
[5] A. Wilson, S. Buolamwini, and T. Gebru, "Predictive Inequity in Object Detection," *arXiv preprint*, arXiv:1902.11097, 2019.
[6] Reddit, "Bias in pedestrian detection AI under low-light conditions," *r/Futurology*, Aug. 2023. Available: https://www.reddit.com/r/Futurology/
[7] "Advanced driver-assistance system," *Wikipedia*, Jul. 2025. [Online]. Available:https://en.wikipedia.org/wiki/Advanced driver-assistance _system
[8] European Commission, "Study on the effectiveness of advanced driver assistance systems," Brussels, Belgium, Rep. No. MOVE/C2/SER/2017263, 2020.
[9] A. Kendall et al., "Learning to drive in a day," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 8248–8254.
[10] C. Chen et al., "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1907–1915.
[11] S. Amershi et al., "Guidelines for human-AI interaction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.
[12] A. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Mask R-CNN based detection under adverse weather conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 0–0.
[14] Y. Li et al., "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," *arXiv preprint arXiv:2002.11297*, 2020.
[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6402–6413.
[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
[17] J. Levinson and S. Thrun, "Robust vehicle localization in urban environments using probabilistic maps," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2010, pp. 4372–4378.

[18] National Transportation Safety Board (NTSB), "Collision between vehicle controlled by developmental automated driving system and pedestrian," Rep. NTSB/HAR-19/03, Nov. 2019.
[19] Reddit, "Radar vs camera performance in fog for AV perception," *r/SelfDrivingCars*, Mar. 2023. [Online]. Available: https://www.reddit. com/r/SelfDrivingCars/
[20] J. Dickmann et al., "Automotive radar: The key technology for autonomous driving: From detection to environmental understanding," in *Proc. IEEE Radar Conf.*, 2016, pp. 1–6.
[21] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?," RAND Corporation, 2016.
[22] M. Hafner et al., "Cooperative and adversarial driver models for testing autonomous vehicles," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 126– 133.
[23] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
[24] L. Ma, F. Zhang, and X. Yang, "Perception safety monitors for autonomous driving: An uncertainty-aware approach," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2021, pp. 159–166.
[25] C. Koopman and P. Koopman, "Safety assurance cases for autonomous vehicle development," *SAE Int. J. Transp. Saf.*, vol. 9, no. 2, pp. 1–14, 2021.
[26] SAE International, "Taxonomy and definitions for terms related to driving automation systems," *SAE J3016*, 2021.
[27] K. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of selfdriving and highly automated vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 90–104, 2016.
[28] J. Koopman, P. Koopman, and M. Wagner, "Remote human supervision of autonomous systems," in *Proc. IEEE Int. Symp. Syst. Eng.*, 2019, pp. 1–8.
[29] M. Kendall and A. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5574–5584.
[30] C. E. Dickerson, "Redundancy and cross-checks in safety-critical automotive computing," in *Proc. SAE World Congr.*, 2020, pp. 1–9.
[31] S. Shalev-Shwartz et al., "On a formal model of safe and scalable selfdriving cars," *arXiv preprint arXiv:1708.06374*, 2017.
[32] T. Buolamwini and G. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness Accountability Transparency*, 2018, pp. 77–91.
[33] Y. Wang et al., "Enhancing autonomous driving perception with synthetic data," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2020, pp. 1–7.
[34] J. Raji et al., "Closing the AI accountability gap," in *Proc. ACM Conf. Fairness Accountability Transparency*, 2020, pp. 33–44.
[35] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness Accountability Transparency*, 2019, pp. 220–229.
[36] A. Beutel et al., "Data decisions and theoretical implications when adversarially de-biasing," in *Proc. Conf. Fairness Accountability Transparency*, 2019, pp. 50–59.
[37] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Transp. Saf.*, vol. 4, no. 1, pp. 15–24, 2016. [38] M. Winner et al., "Virtual testing for autonomous driving," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 207–212.
[38] Z. Yan et al., "A comprehensive survey of digital twinPart 2: Roles of deep learning and blockchain in the industrial internet of things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 4422–4442, 2021.
[39] ISO 21448, "Road vehicles Safety of the intended functionality (SOTIF)," International Organization for Standardization, 2022.
[40] G. Pokharkar, "Scenario-Based Validation for SAE Level 2+ Features Using Simulation-in-the-Loop (SiL) Systems," *International Journal of Innovative Research and Creative Technology*, vol. 11, no. 4, pp. 1–8, Jul. 2025, doi: 10.5281/zenodo.16883284.
[41] E. Thorn, S. Kimmel, and M. Chaka, *A Framework for Automated Driving System Testable Cases and Scenarios*, Report No. DOT HS 812 623, National Highway Traffic Safety Administration, Washington, DC, Sep. 2018.
[42] R. Bishop, "A survey of intelligent vehicle applications worldwide," *IEEE Intell. Veh. Symp.*, pp. 25–30, 2020.
[43] M. Aly, "Real time detection of lane markers in urban streets," *IEEE Intell. Veh. Symp.*, pp. 7–12, 2008.
[44] N. Bojarski et al., "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.
[45] A. Lambert et al., "Over-the-air software updates for connected and autonomous vehicles: Challenges and opportunities," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2103–2116, 2021.
[46] P. Koopman, "A case study of safety-critical embedded software," *IEEE Des. Test*, vol. 29, no. 3, pp. 20–25, 2012.
[47] HLDI, "Real-world benefits of crash avoidance technologies," Highway Loss Data Institute, 2020.
[48] Road & Track, "ADAS systems can change how people drive," 2023.
[49] Euro NCAP, "Assessment protocolHighway Assist," 2022.

[50] UNECE, “UN Regulation No. 157Automated Lane Keeping Systems (ALKS),” 2021.
[51] A. R. P. James et al., “Legal and institutional frameworks for autonomous vehicles in the EU,” *Eur. Transp. Res. Rev.*, vol. 12, 2020.
[52] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
[53] C. Michelmore et al., “Uncertainty quantification for deep learning in autonomous driving,” in *Proc. ICRA*, pp. 208–214, 2020.
[54] W. Samek et al., “Explainable AI: Interpreting, explaining and visualizing deep learning,” Springer, 2019.
[55] S. Vora et al., “Bias in visual AI: Evaluation and mitigation for pedestrian detection,” *arXiv preprint arXiv:2103.03450*, 2021.
[56] Thirunagalingam, A. (2024). Bias Detection and Mitigation in Data Pipelines: Ensuring Fairness and Accuracy in Machine Learning. Available at SSRN 5047605.
[57] Maroju, P. K. (2024). Advancing synergy of computing and artificial intelligence with innovations challenges and future prospects. FMDB Transactions on Sustainable Intelligent Networks, 1(1), 1-14.
[58] L. N. R. Mudunuri, V. M. Aragani, and P. K. Maroju, "Enhancing Cybersecurity in Banking: Best Practices and Solutions for Securing the Digital Supply Chain," Journal of Computational Analysis and Applications, vol. 33, no. 8, pp. 929-936, Sep. 2024.
[59] Settibathini, V. S., Virmani, A., Kuppam, M., S., N., Manikandan, S., & C., E. (2024). Shedding Light on Dataset Influence for More Transparent Machine Learning. In P. Paramasivan, S. Rajest, K. Chinnusamy, R. Regin, & F. John Joseph (Eds.), Explainable AI Applications for Human Behavior Analysis (pp. 33-48). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1355-8.ch003.
[60] Sehrawat, S. K., Dutta, P. K., Bhatia, A. B., & Whig, P. (2024). Predicting Demand in Supply Chain Networks With Quantum Machine Learning Approach. In A. Hassan, P. Bhattacharya, P. Dutta, J. Verma, & N. Kundu (Eds.), *Quantum Computing and Supply Chain Management: A New Era of Optimization* (pp. 33-47). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-4107-0.ch002
[61] S. Panyaram, "Connected Cars, Connected Customers: The Role of AI and ML in Automotive Engagement," International Transactions in Artificial Intelligence, vol. 7, no. 7, pp. 1-15, 2023.
[62] Mohanarajesh Kommineni. (2022/9/30). Discover the Intersection Between AI and Robotics in Developing Autonomous Systems for Use in the Human World and Cloud Computing. International Numeric Journal of Machine Learning and Robots. 6. 1-19. Injmr. - 1