*machines*

MDPI

*Systematic Review*

# A Systematic Review on Risk Management and Enhancing Reliability in Autonomous Vehicles

**Ali Mahmood [1,2,*] and Róbert Szabolcsi [3]**

[1] Doctoral School on Safety and Security Sciences, Óbuda University, 1088 Budapest, Hungary
[2] Systems and Control Engineering Department, Ninevah University, Mosul 41002, Iraq
[3] Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, 1034 Budapest, Hungary; szabolcsi.robert@uni-obuda.hu
* Correspondence: ali.mahmood@uni-obuda.hu

**Abstract**

Autonomous vehicles (AVs) hold the potential to revolutionize transportation by improving safety, operational efficiency, and environmental impact. However, ensuring reliability and safety in real-world conditions remains a major challenge. Based on an in-depth examination of 33 peer-reviewed studies (2015–2025), this systematic review organizes advancements across five key domains: fault detection and diagnosis (FDD), collision avoidance and decision making, system reliability and resilience, validation and verification (V&V), and safety evaluation. It integrates both hardware- and software-level perspectives, with a focus on emerging techniques such as Bayesian behavior prediction, uncertainty-aware control, and set-based fault detection to enhance operational robustness. Despite these advances, this review identifies persistent challenges, including limited cross-layer fault modeling, lack of formal verification for learning-based components, and the scarcity of scenario-driven validation datasets. To address these gaps, this paper proposes future directions such as verifiable machine learning, unified fault propagation models, digital twin-based reliability frameworks, and cyber-physical threat modeling. This review offers a comprehensive reference for developing certifiable, context-aware, and fail-operational autonomous driving systems, contributing to the broader goal of ensuring safe and trustworthy AV deployment.

**Keywords:** Autonomous Vehicles (AVs); safety engineering; Fault Detection and Diagnosis (FDD); collision avoidance; risk assessment; system reliability; Validation and Verification (V&V)

check for updates

Academic Editors: Karim Ahmadi and Seyed Yaser Nabavi Chashmi

Received: 23 June 2025
Revised: 18 July 2025
Accepted: 22 July 2025
Published: 24 July 2025

**Citation:** Mahmood, A.; Szabolcsi, R. A Systematic Review on Risk Management and Enhancing Reliability in Autonomous Vehicles. *Machines* **2025**, *13*, 646. https://doi.org/10.3390/machines13080646

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

By promising increased safety, enhanced mobility, and reduced environmental impact, autonomous vehicles (AVs) are redefining the future of transportation. Designed to operate independently across complex driving scenarios without human oversight, AVs hold great potential. However, despite rapid progress, real-world environments introduce random variables—such as sensor noise, hardware degradation, and complex traffic dynamics—that make ensuring AV safety and reliability especially challenging. The existence of these factors makes it difficult to guarantee reliability and safety in rare-event scenarios [1]. The gap between experimental performance and real-world safety levels is highlighted by high-profile incidents of AVs. Therefore, research into safety-centric methodologies has accelerated due to these concerns, leading to progress across multiple domains of AV system design.

There are several areas of focus, including real-time risk assessment and mitigation strategies [2–4], robust fault detection and diagnosis mechanisms [5–7], and resilient decision making under uncertainty [8,9]. Equally important are system-level architectural approaches that have been designed to ensure reliability during degraded operations [10,11] as well as the frameworks of the rigorous validation and verification for simulated and real-world testing [12].

Due to the multifaceted and interdisciplinary nature of AV safety, a comprehensive review is needed to synthesize technical advancements across perception, planning, control, and systems engineering. By providing a structured and systematic review, this paper aims to fill the gap in current research related to FDD, collision avoidance and decision making, system reliability enhancement, V&V, and performance evaluation.

### *1.1. Background on Autonomous Vehicles*

AVs are systems that rely on integrated technologies from robotics, AI, sensor fusion, and advanced control systems. As defined by the Society of Automotive Engineers (SAE), these systems are categorized into six levels of driving automation ranging from Level 0 (no automation) to Level 5 (full automation). Most of the research and development efforts currently focus on achieving Level 3 (conditional automation) and Level 4 (high automation), in which the vehicle can perform all driving tasks within specific operational design domains (ODDs). Fundamentally, the layered architecture of AVs consists of sensing, perception, localization, decision making, planning, and control. Each layer has its own failure modes, making architectural resilience critical to AV safety [7,13–15]. The perception layer generates an accurate representation of the surrounding environment after integrating data from various sensors such as LiDAR, radar, cameras, and GNSS [13]. In the next step, this information is fed to the planning and decision-making layers, where the current driving context is assessed to select safe, goal-oriented trajectories [3,9,16]. Then, the control layer executes these commands using actuators to ensure smooth and stable motion.

To operate in a real-world environment without any human help makes safety, robustness, and reliability critical for AVs. As AVs must be prepared to handle faults and uncertainties while preserving safety guarantees [13], a growing body of research has emerged to address challenges related to real-time risk modeling [2,4], fault detection and isolation [5,17], and adaptive control under degraded conditions [6]. Recent studies have also explored deep reinforcement learning as a method to improve high-speed cruising performance and adaptive control in AVs. For example, the framework proposed in [18] demonstrates how integrated learning-based strategies can enhance longitudinal stability and responsiveness in dynamic highway conditions.

### *1.2. Motivation for Safety and Reliability Focus*

The deployment of AVs brings challenges in safety beyond those faced in traditional automotive systems. As they work without human oversight, they are solely responsible for perception, decision making, and control in dynamic scenarios. Thus, any malfunction, even minor system faults or misjudgments, can be catastrophic as they directly affect passengers, pedestrians, and other road users [19]. The urgency of improving system safety and resilience has been understood due to several high-profile accidents involving AV prototypes. These incidents indicate that errors arise not only from perception failures or planning mistakes, but also from latent software faults, sensor degradation, or unexpected interactions between subsystems [20]. To address these challenges, safety must be embedded into every layer of system design, starting from perception algorithms and control policies to architectural redundancy and testing protocols. The need for fail-operational behavior has become a central principle in AV design philosophy [6,7,10,21].

As AVs must operate in open-world environments characterized by high uncertainty, incomplete information, and unstructured events, their challenges require real-time risk assessment models [2,3], robust fault detection and diagnosis (FDD) systems [5,17], and resilient decision-making frameworks [8,9]. Validation and verification (V&V) strategies are equally important, as they can effectively test AVs across diverse operational scenarios beyond simple miles driven metrics [2,12]. In this context, ensuring the safety and reliability of autonomous vehicle (AV) systems is not merely a technical requirement; it is also a prerequisite for key societal factors, including societal acceptance, regulatory approval, and the responsible integration of autonomous technologies into public infrastructure. To support this goal, the main objective of this review is to identify, organize, and critically assess the latest research focused on developing safety and reliability in AV systems. Specifically, this paper addresses the following questions: (1) What technical strategies have been proposed to deal with risk, uncertainty, and fault tolerance in AV operations? (2) What are the main limitations and open challenges that prevent the certification and deployment of safe, fail-operational AVs? Answering these questions is essential to guide future development toward more trustworthy and certifiable autonomous systems. While this review focuses on core safety and reliability domains, broader themes such as human–AV interaction and cyber-physical threats are acknowledged as critical but are discussed primarily in the context of future challenges and research gaps.

## 2. Methodology

This review adopted a structured methodology inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure a comprehensive, transparent, and reproducible synthesis of research on autonomous vehicle (AV) safety and reliability. Figure 1 illustrates the process of study selection through a PRISMA flow diagram. In total, 33 peer-reviewed studies were included in the final review corpus.

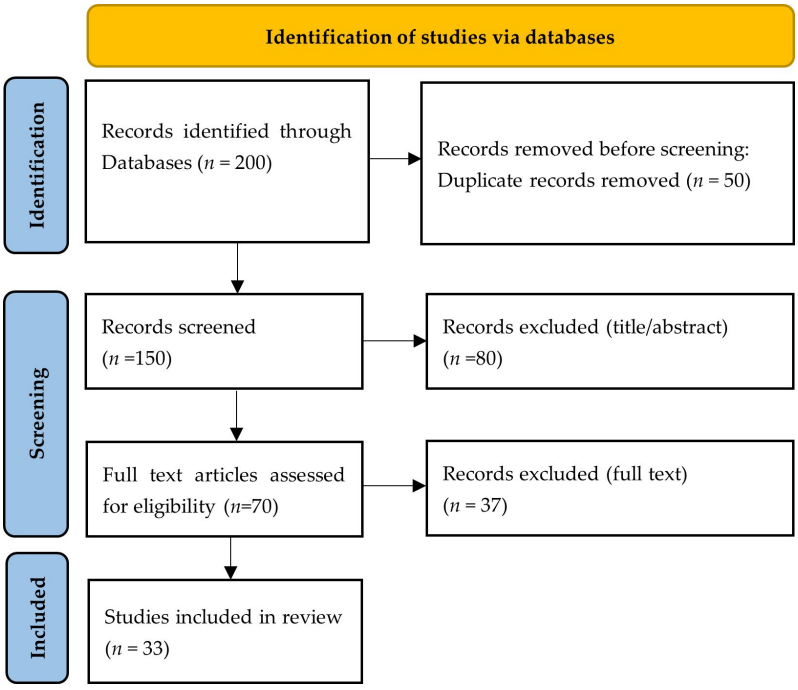


**Figure 1.** PRISMA flow diagram summarizing the literature screening and selection process.

This methodological framework facilitates systematic article selection, promotes the reduction of bias, and ensures coherence and reliability in thematic synthesis. The review process contains three stages: (1) a systematic and replicable search strategy; (2) the defini-

tion of inclusion and exclusion criteria to screen for relevance and quality; and (3) extraction of structured data from the final set of studies.

### 2.1. Search Strategy

A structured literature search was implemented to identify relevant peer-reviewed publications addressing safety, reliability, and decision making in AVs. The search was limited to publications from 2015 to 2025 to focus on recent advancements. Three databases were used: IEEE Xplore, ScienceDirect, and WOS. Journal articles and conference papers were given priority because they present original research or systematic evaluations. An initial set of 200 articles was identified. Several exclusion criteria were applied, starting with the removal of duplicates and irrelevant studies, followed by title and abstract screening, and then full paper review. The final number of papers included was 33, which were downloaded, reviewed in full, and categorized according to their primary focus area.

### 2.2. Inclusion and Exclusion Criteria

A set of predefined inclusion and exclusion criteria was applied during the screening and review process to ensure the relevance and quality of the selected literature.

Inclusion Criteria:

- Studies that have been published in peer-reviewed journals or high-quality conferences;
- Publications that have a focus on safety, reliability, risk assessment, fault handling, decision making, validation, or performance metrics for autonomous vehicles;
- System-level or component-level safety research such as perception, planning, control;
- Papers that propose or evaluate technical solutions (e.g., algorithms, architectures, frameworks);
- Studies published in English between 2015 and 2025.

Exclusion Criteria:

- Studies without technical depth that solely discuss ethical, legal, or social implications;
- Papers that do not focus on autonomous vehicle technologies;
- Studies with limited methodological strength, such as editorials, short abstracts, or opinion pieces;
- Duplicate publications or redundant versions of the same research.

### 2.3. Data Extraction

To systematically gather the relevant information from the selected references, a structured data extraction process has been employed. The aim was to obtain the methodological details and the key contributions that are related to the safety and reliability of autonomous vehicles. The foundation for the structured synthesis presented in the main body of the review has been provided by the thematic categorization, focusing only on the data directly extracted from the texts and figures of the selected papers.

## 3. Fault Detection and Diagnosis (FDD)

As AVs operate in complex and safety-critical environments, resilience against faults in sensors, actuators, and system-level components is essential. The backbone for monitoring, isolating, and mitigating these faults is the Fault Detection and Diagnosis (FDD) system [22]. In this section, the key research contributions are presented to address FDD across various AV subsystems.

### 3.1. Sensor and Perception-Level Fault Detection

This subsection focuses on the used methods to detect faults in AV sensor systems and perception modules, which are essential for maintaining accurate situational awareness.

Fang et al. [5] propose a hybrid FDD system that combines statistical and machine learning techniques. To enhance fault isolation accuracy, this hybrid FDD architecture fuses residual analysis via Kalman filtering with a one-class SVM classifier and decision fusion layers. It enables real-time response to abnormal sensor behavior such as GPS drift. As illustrated in Figure 2, the hybrid FDD architecture supports multi-sensor inputs, feature extraction, and layered fault classification [5].

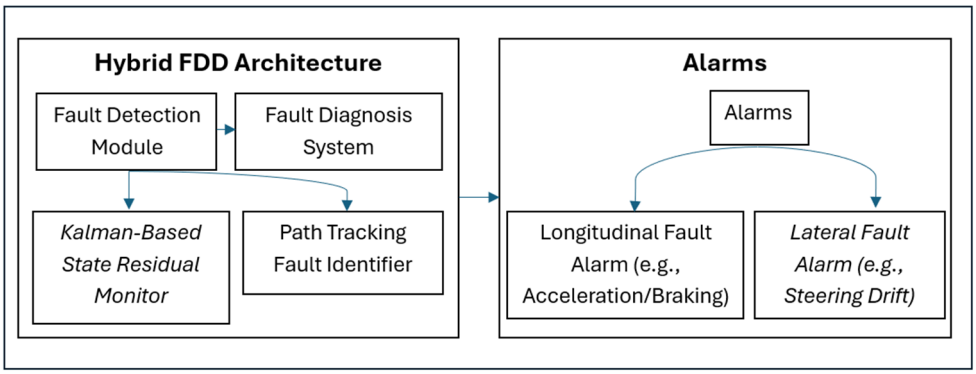


**Figure 2.** Hybrid fault detection and diagnosis system integrating Kalman residuals and machine learning-based classification [5].

By quantifying detection confidence based on object tracking continuity and prediction variance, Liu et al. [17] further advances sensor FDD. The system identifies perceptual faults and avoids false alarms by applying spatiotemporal consistency checks [19,22]. Similar approaches using hybrid soft computing techniques have been applied in mobile robotics. Stefanoni et al. [23] compared FIS and ANN architectures for compensating magnetic localization errors caused by ferromagnetic disturbances, demonstrating robust classification performance under varying environmental noise conditions. In [13], a Fault Detection, Isolation, and Recovery (FDIR) module that fuses GNSS and LIDAR signals was used to address fault-tolerant localization. Then, via dynamic sensor re-weighting, deviations between sensor readings were evaluated, ensuring continuity in vehicle localization even in GPS-denied or partially degraded environments. Compared to purely statistical or ML-based approaches, fusion-based methods like FDIR give better resilience under sensor degradation but often include higher system complexity.

### *3.2. Actuator Faults and Emergency Control*

In this subsection, the focus is on fault detection in actuators and the control strategies AVs use to maintain stability during partial failures. A Sliding Mode Observer (SMO)-based detection scheme was developed by Lee et al. [6], integrated with an emergency fallback control system, and similar efforts have applied H∞ and LQR strategies to achieve optimal suspension performance under uncertain dynamic conditions. The system identifies perceptual faults [23]. When actuator abnormalities are detected, the vehicle transitions into a safe control mode using reduced dynamic envelopes; this ensures stability under partial degradation. A set-based fault detection method leveraging Linear Parameter-Varying (LPV) system modeling is presented in Reference [24]. This method flags deviations that indicate faults without being overly sensitive to model uncertainty by constructing invariant sets for nominal dynamics [9].

For electric AVs, Hashemi et al. [15] focuse on early detection of inter-turn faults in electric drive motors. The system reliably distinguishes between healthy and faulty motor conditions under variable loads using a zero-sequence current and wavelet energy signatures.

### 3.3. Distributed and Networked Fault Detection

This part reviews methods to cooperate fault detection across connected AV fleets, especially in multi-vehicle systems. Cooperative fault detection is a growing necessity as AVs become increasingly connected. Khalil et al. [11] propose a distributed detection and localization system that uses transmissibility functions across a vehicle network. By analyzing the response propagation paths, AV faults are inferred. The effectiveness of this method is particularly evident in platooning or swarm coordination scenarios.

Complementing this, an output-only FDD scheme is introduced in [25], where each vehicle detects faults using only its own output signals. This approach is suitable for large-scale deployment where full system observability is limited.

### 3.4. Fault Injection and Modular Safety Architecture

This subsection discusses techniques to validate FDD systems through simulated faults and highlights modular safety frameworks. A fault injection platform for systematically validating FDD schemes was introduced in Reference [26]. The architecture supports the injection of both hardware-level and software-level faults in simulated environments. The evaluation of the AV systems occurs under fault scenarios, enabling safety performance benchmarking [26].

One of the references at the system architecture level is Reference [10], which proposes an Integrated Modular Safety System (IMSS) that incorporates HAZOP, AV-specific Layer of Protection Analysis (LOPA), and ROS-based supervisory control. This framework enables redundancy management and dynamic fault mitigation.

Across the reviewed FDD approaches, different trade-offs become obvious depending on factors such as system complexity, uncertainty tolerance, and data availability. Hybrid architectures that merge statistical filtering with machine learning techniques (e.g., Kalman + SVM) enable adaptive fault detection across a variety of heterogeneous sensors, though they frequently depend on labeled datasets and can be difficult to interpret. In contrast, model-based observers like a Sliding Mode Observer (SMO) and Linear Parameter Varying (LPV) methods offer mathematically grounded performance guarantees and are well-suited to structured dynamic environments, but they are also sensitive to modeling errors and need precise system identification. Set-based fault detection methods are efficient in dealing with uncertainty and developing robustness, yet they tend to be demanding computationally, which limits their practicality for real-time implementation. Networked FDD systems suggest scalability for connected vehicle fleets but may face issues with synchronization and communication reliability. Together, these comparisons highlight the importance of aligning FDD strategies with exact operational needs such as actuation in high-speed, degraded localization or limited observability as no single method can address all scenarios optimally. Combining multiple complementary approaches within modular safety architectures, such as IMSS, appears to be a promising solution for achieving resilient and certifiable fault management in autonomous vehicles.

Across different AV contexts, these FDD methods vary in terms of effectiveness. In urban settings where traffic and occlusions are unpredictable, hybrid learning-based systems have better handling for sensor anomalies, while model-based observers struggle with high modeling errors. On structured highways, reliable fault isolation has been offered by techniques such as SMO and LPV. For sensor-degraded, fusion-based conditions, FDIR provides better localization than single-sensor models.

## 4. Collision Avoidance and Decision Making

To avoid collisions while maintaining safe, efficient, and lawful navigation, AVs must continuously make real-time decisions. This involves evaluating risks, predicting trajectories, and planning safe maneuvers under uncertainty [9].

### 4.1. Risk-Informed Collision Avoidance Strategies

This subsection presents strategies that give permission to AVs to assess and respond to potential collision risks using environmental and behavioral cues. A collision risk assessment framework was developed by Reference [2], which integrates environmental awareness, trajectory prediction, and fuzzy logic. To help AVs make proactive avoidance decisions, this system calculates the probability of a collision using multi-agent predicted paths. By quantifying spatiotemporal threats in structured urban environments, this layered risk estimation supports decision making.

Lin and Tsukada [4] introduce a path planning algorithm based on an adaptive potential field (APF). It integrates risk components, including relative velocity, proximity, and road constraints, into the potential field computation. Using this approach, AVs can generate dynamically safe and smooth trajectories, even in complex driving scenarios. A decision-making system that considers the severity of potential collisions and selects the least risky maneuver using a scenario-based assessment is presented in [9]. Leng et al. [27] introduce a method that improves responsiveness during rapid evasive maneuvers while balancing safety and efficiency. This is done by introducing a hierarchical reinforcement learning framework where risk-aware value functions guide navigation policies.

In addition to the strategies that have been discussed, many external and operational factors influence the effectiveness of risk management in autonomous vehicles. These include factors such as adverse weather, road conditions, sensor degradation, and the unpredictability of human behavior. Robust control strategies, such as the work by Zhang et al. [28], have presented promise in improving path tracking under uncertainty, thereby enhancing overall system safety.

### 4.2. Scenario-Based and Context-Aware Planning

This section includes planning methods that regulate AV behavior based on road types, traffic complexity, and context-specific risks. Machine learning has been applied to dynamic risk assessment in diverse driving scenes. The system evaluates environmental, behavioral, and traffic-based features based on datasets capturing driving patterns to generate risk scores, which are then used for downstream decision making [8]. Recent work by Yang et al. [29] further applies risk-aware reinforcement learning to safely manage intersection scenarios, demonstrating improved maneuver planning during high-risk merging and yielding tasks. This gives AVs the ability to find optimal actions based on predicted scene risks, such as yielding, lane changing, or slowing.

By investigating risk variability across driving contexts, it has been identified that intersections and rural highways present distinct challenges [12]. Thus, their analysis highlights the importance of adjusting AV policies according to road geometry, traffic density, and maneuver types. Their findings inform people of the need for environment-specific decision rules.

By Sheikh and Peng [16], merging behavior was enhanced using real-time collision risk estimation at highway on-ramps. AVs reduce conflict with mainline traffic by calculating dynamic risk values and adjusting velocity and merging strategy. The decision loop is shown in Figure 3, where risk-informed control parameters influence the lane entry path [17].

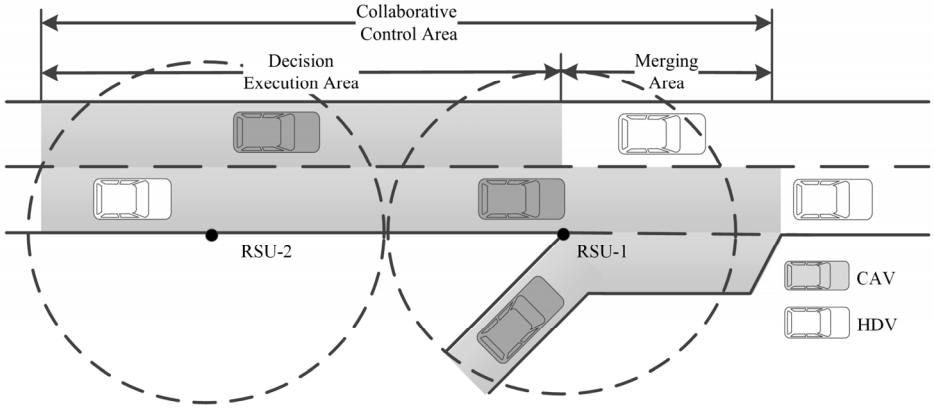


**Figure 3.** Highway on-ramp merging system integrating dynamic risk modeling into decision making and control [17].

A Bayesian behavior prediction model for multi-agent environments has been used to enhance AV decision making in ambiguous interactions. This model has the ability to learn from real-world datasets to anticipate complex maneuvers such as merging or yielding, enabling context-sensitive, probabilistically robust planning under behavioral uncertainty [30,31].

### *4.3. Uncertainty-Aware Decision and Safety Envelopes*

In this section, the emphasis is on frameworks of decision making that explicitly manage uncertainty to ensure safe trajectory selection. Probabilistic safety envelopes have been introduced for decision making under perception uncertainty, where these envelopes represent confidence bounds around obstacles and trajectory predictions to maintain a probabilistically safe distance [22]. The framework incorporates aleatoric uncertainty into the motion planner while preventing aggressive behaviors in ambiguous scenes.

In addition, an adaptive potential field-based planner was introduced in [3] and was used to reshape navigation vectors in real time based on surrounding risk, enabling smooth yet safe avoidance decisions. The adaptive safety zone expansion based on uncertainty levels is illustrated in Figure 4 [3]. These methods expand risk sensitivity in uncertain scenes, though their performance still relies on reliable uncertainty quantification from upstream modules.

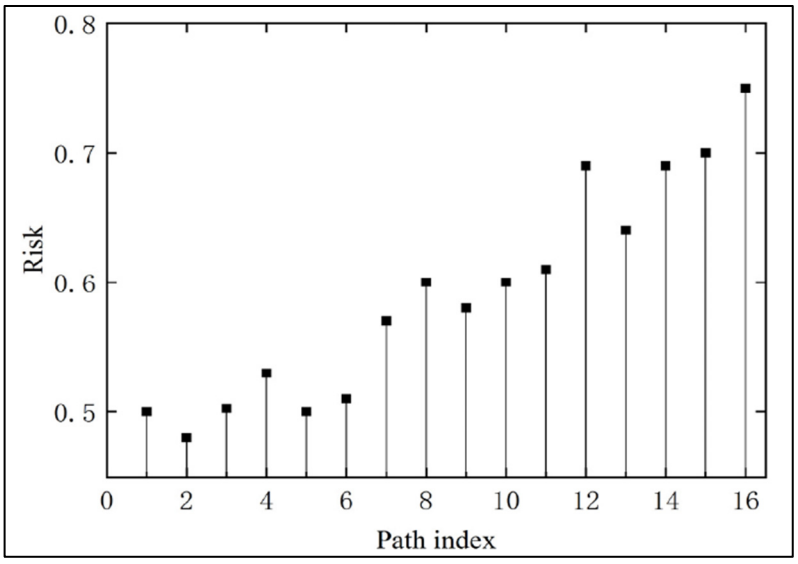


**Figure 4.** Adaptive potential field contours guiding decision making under risk awareness and environmental constraints [3].

Collision avoidance was integrated with a modular safety architecture in Reference [10]. This system uses Safety of the Intended Functionality (SOTIF) principles and real-time risk flags to manipulate trajectory planning layers. AVs react safely to detected risks, even under degraded perception or control. Finally, a comprehensive review of uncertainty quantification (UQ) methods was provided by [32], identifying their importance for AV reliability. Under the Probabilistic Risk Assessment for Software Engineering (PRASE), Bayesian ensemble and deterministic approaches have been evaluated, highlighting their trade-offs in robustness, accuracy, and efficiency [15]. This work highlights the need for a UQ-aware decision-making pipeline, especially in scenarios such as perception degradation or out-of-distribution conditions.

The decision-making methods reviewed differ in a significant way in terms of complexity, adaptability, and ease of interpretation. Rule-based methods and potential field approaches, such as APF, are efficient in terms of computation but can express difficulties when operating in highly dynamic or ambiguous environments. Learning-based strategies, containing reinforcement learning and Bayesian behavior models, offer better adaptability in unstructured settings; however, they frequently rely on large datasets and pose challenges related to transparency and formal verification. Scenario-based planners afford a middle ground by combining both structure and flexibility, but they rely on comprehensive scenario libraries to prevent coverage gaps. These distinctions highlight the inherent trade-offs between real-time responsiveness, system robustness, and the ability to guarantee safety through formal methods. As a result, hybrid pipelines that integrate interpretable models with learning-based components are emerging as a promising direction to achieve both practical performance and reliable decision making in autonomous systems.

Beyond these general trade-offs, the effectiveness of these methods also relies on operational context. In urban traffic, Bayesian models and reinforcement learning result in better ambiguity and interaction. Rule-based and potential field methods perform efficiently on structured highways. In occluded zones, safety envelopes help mitigate risk, while scenario-based planners excel in dynamic tasks such as merging. These differences reinforce the case for hybrid frameworks tailored to specific driving scenarios.

## 5. System Reliability and Resilience

As AVs are deployed in dynamic, real-world environments, it is critical to ensure reliable, fail-operational performance. In case of fault detection or collision avoidance, AVs must be designed in a way that maintains safe operation across a spectrum of rare events, hardware degradation, and contextual uncertainties. In this section, recent approaches are synthesized with the aim of enhancing long-term system reliability and architectural resilience.

### *5.1. Modular Safety Architectures and Functional Redundancy*

This subsection outlines architectural designs that support continued AV operation under subsystem failures through built-in redundancy. For AVs to maintain essential functionality even during subsystem faults, it is required to have safe and reliable autonomy [7]. An Integrated Modular Safety System (IMSS) was developed by [10].

The IMSS framework has a layered safety control structure. It combines overarching mission execution with detailed safety enforcement at a lower level. To monitor functional blocks, Safety Programmable Logic Controllers (PLCs), supervisory watchdogs, and diagnostic loops are employed; this includes perception, planning, and actuation, allowing fault isolation before faults cascade across the system. To allow continued operation under partial failures, each safety-critical component is designed with backup modes or redundant counterparts.

The system architecture implemented on the nUWAy autonomous shuttle features both traditional safety components and AI-based modules for dynamic hazard assessment. Middleware interfaces are used to coordinate communication across safety and mission layers, ensuring real-time responses to system health indicators and external conditions [10,14].

### *5.2. Emergency Control and Adaptive Resilience*

This section evaluates adaptive control techniques to ensure vehicle stability and safety during fault conditions. Adaptive emergency control mechanisms were developed by Lee et al. [6]. These mechanisms maintain vehicle stability and trajectory compliance under partial failures. When the system detects anomalous behavior in actuators using a Sliding Mode Observer (SMO), it immediately engages an emergency control layer that modifies control limits and response gains. By ensuring the AV can perform minimal-risk maneuvers, this approach supports fail-operational behavior. This includes controlled braking or lateral stabilization, even in the presence of steering or propulsion faults [6]. By actively supporting degraded performance rather than immediate shutdown, reliability beyond fault detection is extended. While emergency fallbacks support controllers' real-time safety, they must be carefully tuned to prevent overreaction, especially when merged with adaptive strategies in uncertain conditions.

### *5.3. Context-Aware Reliability Across Operational Domains*

This part examines AV reliability frameworks adapted to varying road environments, driving scenarios, and operational domains. Internal component resilience is not the only factor used to determine reliability; the way the system responds to environmental changes is also crucial. This issue was addressed by Shetty et al. [12], where the risk and performance of AVs were analyzed across diverse operational contexts, including intersections, highways, and urban roads [33]. In this study, risk distribution is considered highly context-sensitive, while AV reliability frameworks incorporate environment-specific behavior adjustment. According to the findings of this study, AV systems must not only survive component faults but also adapt their operational strategies. This includes maneuver aggressiveness, lane selection, and safety margin selection. This adaptation through a context-aware approach enhances reliability in long-term deployments [34].

Several strategies intended at developing system reliability bring different yet complementary advantages. Modular architectures, such as IMSS, propose a well-organized method of fault containment and redundancy management; however, they frequently need significant effort in system integration and may rely on the presence of hardware redundancy. Adaptive emergency control methods are capable of handling faults in real time but are extremely dependent on accurate fault detection mechanisms and tuning of control parameters. Context-aware dependability frameworks support adaptation to changes in environmental conditions but depend on comprehensive contextual modeling and the ability to interpret real-time data efficiently. Each of these approaches targets a distinct layer of resilience, whether at the hardware level, within control systems, or in behavioral decision making. Mixing these layers into cohesive and certifiable safety architectures is still both a significant challenge and a promising direction for advancing the design of autonomous vehicles.

Outside technical distinctions, these strategies also change in relevance across deployment contexts. The architectures of modular safety are ideal for high-assurance highway scenarios, while adaptive emergency controls have flexibility in urban areas with different disruptions. In contrast, context-aware frameworks are essential in diverse environments like intersections, rural roads, or dense traffic. Here, the operating risks are highly variable.

## 6. Validation and Verification (V&V)

The foundation of safety assurance in autonomous vehicles (AVs) is established through rigorous validation and verification (V&V) processes. However, the complexity of AV perception, planning, and control modules—combined with the need to operate in dynamic and unpredictable environments—presents significant challenges, making comprehensive safety coverage difficult to achieve. In this section, recent approaches and frameworks that extend V&V beyond traditional performance metrics will be highlighted. These include systematic methods grounded in established safety standards, aiming to ensure robust and certifiable AV operation across diverse operational domains.

### *6.1. Standards-Based Safety Engineering and Modular Assurance*

This subsection goes through formal methods and standards that guide the design and validation of safety-critical AV architectures. An Integrated Modular Safety System (IMSS) architecture, developed in alignment with ISO 26262, ISO/PAS 21448 (SOTIF) and emerging Safety of the Intended Artificial Intelligence (SOTAI) framework, which is under development, was introduced in [10,35,36]. To ensure functional safety compliance across the perception, planning, and actuation layers, the IMSS incorporates modular fault containment strategies, hazard propagation control mechanisms, and Robot Operating System (ROS)-based supervisory diagnostics [14]. In particular, the safety-oriented design includes the following key components:

- For systematic hazard identification: HAZOP (Hazard and Operability) analysis;
- For risk quantification: AV-specific LOPA (Layer of Protection Analysis);
- To detect and mitigate residual faults: diagnostic redundancy.

By formalizing safety case development and enforcing traceability between system design requirements and corresponding validation procedures, contemporary verification and validation (V&V) methods extend beyond traditional simulation-driven testing approaches [7]. In this context, Saulaiman et al. [37] proposed an automated Threat Analysis and Risk Assessment (TARA) framework, which systematically generates attack graphs to evaluate cybersecurity vulnerabilities in autonomous vehicle (AV) architectures. Their method adheres to the ISO/SAE 21434 cybersecurity standard [38], offering a structured and repeatable process for identifying threat propagation paths and quantifying associated risks. Furthermore, the framework facilitates the validation of implemented countermeasures by assessing their effectiveness in mitigating identified threats, thereby enhancing the overall safety assurance argument for AV systems operating in connected and adversarial environments.

### *6.2. Scenario-Based Evaluation and Contextual Safety Metrics*

In this section, the focus is on assessment approaches that test AV performance in detailed, realistic driving scenarios instead of broad aggregates. For assessing the safety performance of autonomous vehicles (AVs), Shetty et al. [12] critically examine the limitations of commonly used aggregate safety metrics, such as “miles per crash” or “disengagements per 1000 km.” While these metrics are easily quantifiable and often used in public reporting, the authors argue that they are insufficiently granular to capture the nuanced and high-consequence risks inherent in specific operational contexts. Such aggregate indicators may obscure critical safety vulnerabilities that manifest during rare or complex driving scenarios—such as unprotected left turns, multi-agent interactions at intersections, or dynamic changes in urban environments. The study emphasizes the need for context-aware evaluation frameworks that can better reflect situational risk exposure and performance, thereby providing a more rigorous basis for the validation and verification of AV safety across diverse real-world conditions.

To address the limitations of aggregate safety metrics, a scenario-based evaluation framework has been proposed to systematically test AV behavior across well-defined and safety-critical driving situations. These include scenarios such as unprotected left turns at unsignalized intersections, multi-lane highway merges, and occluded pedestrian crossings [21].

The importance of developing test datasets and simulation environments has been emphasized to realistically represent the diversity of road conditions, vehicle behaviors, and environmental factors found in real-world driving. Scenario variability should be a core feature of AV validation pipelines, including maneuver type, interaction complexity, and traffic density [1,19].

The crash probability associated with different maneuvers varies significantly, as shown in Table 1. Due to their frequent occurrence, left-turn maneuvers exhibit the highest probability of a crash, highlighting the disproportionate risk associated with certain behaviors. Lane keeping results in the highest number of recorded crashes, primarily due to its high prevalence during driving [12]. Both crash probabilities and counts are derived from visual interpretation of the plotted data, providing approximate but informative comparisons across maneuver types.

**Table 1.** Estimated crash probabilities and crash counts across driving maneuvers [12].

| Maneuver | Estimated Crash Probability | Crash Count |
|---|---|---|
| Left Turn | $8 \times 10^{-7}$ | 115 |
| Right Turn | $9 \times 10^{-8}$ | 40 |
| Crossing Straight | $5.5 \times 10^{-8}$ | 180 |
| Lane Change | $3.2 \times 10^{-8}$ | 20 |
| Lane Keeping | $6.5 \times 10^{-8}$ | 202 |

Standards-based validation and verification (V&V) frameworks, containing those grounded in ISO 26262 and SOTIF, offer structured traceability and ensure regulatory compliance; however, they fall short when it comes to addressing the dynamic behavior of components driven by machine learning. In contrast, scenario-based evaluation methods are efficient in capturing real-world variability and low-probability, high-impact events, yet they often lack the formal guarantees needed for exhaustive system coverage.

Cybersecurity-focused methods, such as TARA, broaden the validation landscape by incorporating threat modeling and risk assessment, but their integration with functional safety processes stays a challenge. Each of these methodologies supports a distinct layer of the assurance framework—whether regulatory, behavioral, or adversarial—highlighting the necessity for integrated strategies that unify these methods to achieve robust and supportable safety across both conventional and AI-based AV subsystems.

The suitability of V&V strategies also depends on deployment context. Standards-based methods (e.g., ISO 26262, SOTIF) work well for structured environments with deterministic behavior, such as highways or fixed-route AVs. In contrast, scenario-based evaluation is critical in urban or mixed-traffic settings, where unpredictable events require richer behavioral testing. Cybersecurity-focused approaches like TARA are essential in connected AV deployments, where network vulnerabilities could trigger safety failures. An effective V&V framework must integrate these methods based on the operational and threat landscape of the AV.

## 7. Safety Metrics and Evaluation

For validating the safety claims of AVs and benchmarking their behavior under various operational contexts, robust performance evaluation is essential. Traditional metrics such

as crash rates per mile or disengagements over distance offer limited insight into the system's capability to manage rare and complex scenarios. In this section, the approaches for defining and measuring AV safety and performance will be presented.

### 7.1. Context-Aware Safety Metrics

This subsection analyses how the safety metrics can be tailored to reflect the specific risk conditions of different driving contexts and maneuvers. The limitations of using broad and aggregate metrics have been highlighted in Reference [12]. These metrics include overall crash rates or disengagements per mile. The varying levels of risk across specific driving scenarios, such as unprotected turns, have not been adequately captured using high-level statistics. Without contextual segmentation, dangerous behaviors may go unnoticed, leading to incomplete safety profiles [19].

A contextualized safety assessment framework has been developed to evaluate AV decisions and outcomes based on distinct operational scenarios and driving conditions, rather than treating all maneuvers as equally risky. Scenario-based metrics include:

- Crash probabilities (e.g., left-turns, right-turns);
- Time-to-collision (based on maneuver class and speed);
- Exposure-adjusted crash likelihood for different road environments (e.g., urban, sub-urban, highway).

These metrics have been utilized by developers and researchers to identify areas where AVs are most at risk, enabling focused safety enhancements in the most critical regions. This is especially important for verifying the performance of decision-making algorithms and control systems in challenging or high-risk situations.

The probabilities of crash risk vary widely among different maneuver types. Although lane keeping accounts for the highest number of total crashes due to its frequency, left-turn maneuvers exhibit a significantly higher crash probability, indicating an elevated risk despite fewer occurrences [12].

Moreover, Yu et al. [39] emphasize the necessity of refined safety metrics by highlighting the role of Uncertainty Quantification (UQ) in supporting the ISO 21448 (SOTIF) framework. Specifically, UQ supports both levels of safety acceptance standards by using robustness, accuracy, and efficiency metrics. This enables AVs to assess and manage uncertain conditions. These insights facilitate the application of scenario-based safety assessments and enhance trust in AV decisions under unclear or uncertain driving situations.

### 7.2. Model Evaluation and Control Benchmarking

In this subsection, the focus is on assessing models of AV control using statistical and information-theoretic benchmarks to assess tracking performance and reliability. To accurately evaluate the tracking performance of longitudinal control models in AVs, Lyu et al. [31] propose a benchmarking approach that uses multiple statistical and information-theoretic metrics. These include:

- Mean absolute error (MAE) and Root Mean Square Error (RMSE) for quantifying trajectory prediction accuracy;
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for assessing model complexity and generalization.

This benchmarking was applied to four widely used car-following models: the Optimal Velocity Model (OVM), Full Velocity Difference Model (FVD), Intelligent Driver Model (IDM), and the newly proposed Perceived Risk Field Model (PRFM). According to Tan et al. [33], among the four car-following models evaluated, the PRFM model demonstrated the best performance in terms of prediction accuracy, achieving the lowest mean absolute error (MAE) and mean absolute percentage error (MAPE). Although the IDM

model yielded slightly better AIC and BIC values, indicating a simpler model structure, the PRFM provided more realistic and context-sensitive risk modeling. This trade-off is especially valuable for autonomous vehicle safety systems operating under diverse and dynamic driving conditions.

These benchmarking efforts have been further elaborated by Reference [34], which promotes the use of standardized evaluation frameworks for the quantification of uncertainty. Their proposed PRASE framework provides a structured approach to compare safety-related models. By incorporating both predictive accuracy and uncertainty management, this structured evaluation supports aligning control benchmarking with broader safety goals.

The assessment metrics and benchmarking methods reviewed differ significantly in both their emphasis and analytical depth. Aggregate performance indicators, although straightforward to calculate and communicate, frequently overlook critical scenario-specific risk patterns essential for thorough safety validation. In contrast, context-aware metrics enhance the ability to diagnose system behavior but depend on large, behaviorally diverse datasets and detailed scenario annotations. Statistical measures such as mean absolute error (MAE) and Akaike Information Criterion (AIC) offer valuable insights into model precision and complexity, yet they do not inherently capture aspects directly tied to safety assurance. Uncertainty-aware evaluation frameworks such as PRASE address this limitation by integrating predictive accuracy with robustness evaluation under changing conditions. To support substantiated and operationally valid safety claims, an effective evaluation strategy should integrate context-sensitive risk indicators, statistical model performance metrics, and uncertainty quantification methods.

The relevance of safety metrics depends on the AV's operational domain. Aggregate metrics (e.g., disengagements per mile) may suffice for long-haul highway testing but lack granularity in urban scenarios with complex interactions. Context-aware metrics like time-to-collision or maneuver-specific crash probability are essential in dense traffic or intersection-heavy areas. Similarly, uncertainty-aware benchmarks become crucial in environments prone to occlusions or sensor noise. Thus, safety evaluation frameworks must adapt to the specific risk profile and complexity of the deployment context.

## 8. Research Gaps

To achieve certifiable, real-world AV systems, the literature reveals several profound and interconnected gaps that must be addressed despite the extensive advances in each thematic area. These gaps have been organized by cross-cutting challenges and domain-specific deficiencies:

### *8.1. Lack of Cross-Layer Fault Propagation Modeling*

Despite the existing FDD mechanisms detecting localized faults, they do not provide frameworks for modeling the propagation of such faults through AV subsystems. Therefore, probabilistic models (such as Bayesian networks) are needed to map the interdependence between perception, decision making, and control, allowing for more general diagnostics and mitigation. This is mainly because of the absence of united architectures or tools that define dynamic interactions across AV layers. As a result, cascading faults can be undetected, increasing the risk of system-wide failure under runtime uncertainty.

### *8.2. Insufficient Robustness to Distributional Shift and Adversarial Perturbations*

Most AV decision-making algorithms rely on assumptions of stable data distributions. However, out-of-distribution (OOD) inputs and adversarial attacks have been observed in real-world deployments. Therefore, robust OOD detection methods and fallback strategies

are required to maintain safety in novel or corrupted input scenarios. Although some OOD recognition and adversarial training methods occur, they are frequently limited by computational complexity or inadequate generalization. In practice, AVs remain vulnerable to unpredicted conditions, which can lead to degraded or unsafe behavior.

### *8.3. Limited Formal Verification of Learning-Based Modules*

Though central to modern AV perception and control, deep learning components lack formal guarantees. This necessitates novel runtime assurance techniques, such as shielding and SMT-based formal verification, to bridge the gap between empirical accuracy and provable safety. However, these methods are usually limited to simplified models or small-scale networks and do not scale to the complexity of full AV systems. This keeps a critical gap in certifying AI-driven components under formal safety standards.

### *8.4. Neglect of Long-Term Reliability Under Wear and Aging*

Despite current reliability frameworks emphasizing fault-tolerant design, they often ignore component degradation over time. To manage gradual performance drift in long-term deployments, the use of digital twin-based prognostics and adaptive recalibration schemes is essential. However, applying these models requires high-fidelity sensor data over extended periods, which is hardly available in existing deployments. Without this, AVs may face decreased performance or unobserved degradation over time.

### *8.5. Absence of Rich, Scenario-Based Validation Datasets*

The lack of comprehensive datasets constrains the validation and verification (V&V) processes, as these datasets must include near-misses, corner cases, and adverse conditions. Therefore, establishing open repositories with annotated, high-fidelity logs is necessary to support reproducibility and safety benchmarking. Most existing datasets emphasize normal driving, which fails to test the performance of AVs in high-risk or rare-event scenarios. This limits both training robustness and accuracy evaluation for safety-critical behavior.

### *8.6. Underdeveloped Human–AV Interaction Models*

Without models that ensure transparent, predictable, and interpretable interactions with human actors, technical safety measures alone are insufficient. This gap emphasizes the need for both trust-calibrated interfaces and human behavior prediction models. In this area, research is still emerging, and limited AV systems incorporate real-time modeling of pedestrian intent or cooperative negotiation in traffic. This reduces AVs' ability to communicate intent in a clear manner or respond appropriately in shared environments.

### *8.7. Insufficient Cyber-Physical Threat Modeling*

Due to the lack of integrated cyber-physical threat modeling, the ability to identify and mitigate security-driven failure modes is limited. Currently, existing frameworks rarely incorporate cybersecurity standards, such as ISO/SAE 21434, into safety engineering processes or threat analysis methods like TARA (Threat Analysis and Risk Assessment). This creates a blind spot for cyberattacks that exploit vulnerabilities in perception or control systems, thereby undermining the overall resilience of autonomous vehicles.

## 9. Future Directions

The following section offers a roadmap for the safe and reliable integration of AVs based on the identified gaps.

### 9.1. Unified Cross-Layer Fault Models

To represent subsystem interactions and fault chains, Bayesian or factor graph-based models need to be developed. Such models can be applied to diagnosis and mitigation in real time, utilizing automated extraction from AV middleware configuration (e.g., ROS, AUTOSAR). One of the technical challenges is accurately modeling interdependence across perception, planning, and control without overcomplicating the system state. While ROS-based platforms permit modular access to runtime signals, integrating fault propagation logic requires standardized interfaces. Current Bayesian network tools (e.g., PyBN) and dynamic model libraries have yet to scale to full-system AV applications.

### 9.2. Distributional Robustness and Adversarial Resilience

There is a need to integrate uncertainty-aware perception modules using Bayesian deep ensembles with hierarchical planners that trigger fail-safe behaviors upon detecting low confidence or adversarial inputs and to establish standardized adversarial testing benchmarks to validate system robustness. Implementation requires detecting shifts in input distributions in real time, which can be accomplished using confidence thresholds or statistical change detection. While tools exist for adversarial testing in simulation (e.g., CARLA adversarial toolkits), their coverage of physical world conditions are still limited.

### 9.3. Verifiable Learning-Based Components

To validate neural network outputs, formal verification methods, including SMT solvers and interval analysis, must be employed to generate runtime shields that detect constraint breaches and enforce safe operations without retraining, integrating these safeguards within AV middleware layers. Efforts like Reluplex or VeriNet show progress, but they are limited to small networks or limited layers. Inserting these tools into real-time decision pipelines is still an open challenge due to latency and scalability issues.

### 9.4. Prognostics-Driven Reliability Management

There is a need to employ digital twins in conjunction with machine learning-based RUL models to forecast hardware deterioration and deploy real-time recalibration protocols informed by sensor diagnostics to forestall failures in perception and control. Applying these systems requires long-term operational data and robust models that account for nonlinear degradation. While digital twin platforms (e.g., Siemens' MindSphere) are emerging, they are hardly tailored to autonomous systems.

### 9.5. Scenario-Rich Dataset Infrastructure

Standardized, multi-source repositories containing pre-crash, near-miss, and edge-case driving datasets need to be established. Comprehensive metadata schemas and automated edge-case mining tools need to be integrated to enable scalable, reproducible safety analysis and model training. This will require cooperation across industry and academia, along with privacy-compliant data sharing agreements. Current datasets like Waymo Open Dataset or nuScenes are valuable but lack high-frequency rare-event logging or cross-sensor redundancy.

### 9.6. Trust-Centered Human–AV Interaction

Intent recognition models for pedestrians and drivers need to be enhanced using multimodal cues. External Human–Machine Interfaces (HMIs) need to be developed to communicate AV intent transparently and empirically assess their effects on trust, compliance, and interaction safety. Field-tested keys remain limited, and many proposed HMIs do not have standardized assessment protocols. More large-scale user studies are required to evaluate trust dynamics and behavioral adaptation in real-world interactions.

### 9.7. Integrated Cyber-Physical Adversarial Resilience

To enhance system-level resilience, the interplay between cybersecurity and functional safety must be systematically addressed. This involves integrating TARA-based threat modeling into the system design process, alongside the development of runtime intrusion detection systems capable of correlating cyber events with physical anomalies. Furthermore, adversarial conditions should be simulated through testing pipelines that include sensor spoofing and data manipulation to evaluate system robustness. Autonomous vehicle (AV) architectures can adopt a co-assurance strategy by treating cyber and physical threats as interlinked, thereby strengthening protection across both software and hardware domains.

## 10. Conclusions

This systematic review has united recent research developments aimed at strengthening the safety, reliability, and decision-making capabilities of autonomous vehicles (AVs). The synthesis has been structured around five critical thematic areas: fault detection and diagnosis (FDD), collision avoidance and decision making, system reliability enhancement, validation and verification (V&V), and safety performance evaluation.

The analysis of FDD approaches illustrates a shift towards hybrid models that integrate traditional analytical methods with data-driven techniques to enable real-time fault isolation and mitigation across sensing, actuation, and control layers. Concurrently, decision-making frameworks have become increasingly risk-aware and uncertainty-resilient, applying probabilistic planning, scenario prediction, and adaptive policies to enhance responsiveness in complex environments. In terms of system resilience, layered and modular architectures have been shown to facilitate graceful degradation and fail-operational performance during partial failures. Validation and verification (V&V) efforts have moved toward scenario-based safety assessment, moving beyond aggregate statistics to context-sensitive testing frameworks capable of capturing critical edge cases. Additionally, this review highlights the growing importance of domain-specific safety metrics and uncertainty quantification methods in measuring AV performance accurately across diverse operational scenarios. Despite the progress, several research gaps remain, including the limited formal assurance of machine learning-based components, insufficient modeling of inter-subsystem fault propagation, inadequate datasets capturing rare or safety-critical scenarios, and underdeveloped human–AV interaction frameworks.

The future directions proposed in this review directly address these challenges through solutions such as verifiable machine learning, cross-layer fault modeling, uncertainty-aware control, and integrated cybersecurity strategies. These components are essential to meet formal safety assurance requirements and enable certifiable AV behavior under diverse operational conditions. Addressing these challenges will require integrated approaches that incorporate formal verification, distributionally robust learning architectures, cross-layer fault modeling, and context-aware interaction protocols.

Despite this review study involving an extensive range of safety and reliability methods, their maturity levels vary. Some of the approaches, such as modular safety architectures, have aligned with ISO standards, and fault detection by Kalman filtering or SMO have demonstrated deployment in real-world AV systems. However, techniques such as Bayesian behavior prediction, verifiable learning components, or digital twins are still at the experimental or simulation stage. Analyzing this distinction improves identification of which methods are ready for industry adoption and which still require validation before field deployment.

In conclusion, advancing AV safety and reliability necessitates a multi-disciplinary, system-level perspective that merges formal engineering rigor with adaptive intelligence and real-world validation. This synthesis reinforces the central goal of achieving context-

aware, fail-operational, and certifiable autonomy. Such integration is essential to bridge the gap between prototype performance and deployable autonomy, supporting the development of certifiable and trustworthy autonomous mobility systems.

**Author Contributions:** Conceptualization, A.M.; methodology, A.M.; software, A.M.; formal analysis, A.M.; resources, A.M.; writing—original draft preparation, A.M.; validation, R.S.; writing—review and editing, R.S.; supervision, R.S.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Óbuda University, 1034 Budapest, Bécsi út 96/b, Hungary.

**Data Availability Statement:** The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Behzadan, V.; Munir, A. Adversarial Reinforcement Learning Framework for Benchmarking Collision Avoidance Mechanisms in Autonomous Vehicles. *IEEE Intell. Transp. Syst. Mag.* **2021**, *13*, 236–241. [CrossRef]
2. Katrakazas, C.; Quddus, M.; Chen, W.H. A new integrated collision risk assessment methodology for autonomous vehicles. *Accid. Anal. Prev.* **2019**, *127*, 61–79. [CrossRef] [PubMed]
3. Yang, W.; Li, C.; Zhou, Y. A Path Planning Method for Autonomous Vehicles Based on Risk Assessment. *World Electr. Veh. J.* **2022**, *13*, 234. [CrossRef]
4. Lin, P.; Tsukada, M. Adaptive Potential Field with Collision Avoidance for Connected Autonomous Vehicles. In Proceedings of the 2022 13th Asian Control Conference (ASCC), Jeju, Republic of Korea, 4–7 May 2022; pp. 2251–2256. [CrossRef]
5. Fang, Y.; Min, H.; Wang, W.; Xu, Z.; Zhao, X. A Fault Detection and Diagnosis System for Autonomous Vehicles Based on Hybrid Approaches. *IEEE Sens. J.* **2020**, *20*, 9359–9371. [CrossRef]
6. Lee, J.; Oh, K.; Yoon, Y.; Song, T.; Lee, T.; Yi, K. Adaptive Fault Detection and Emergency Control of Autonomous Vehicles for Fail-Safe Systems Using a Sliding Mode Approach. *IEEE Access* **2022**, *10*, 27863–27880. [CrossRef]
7. Iturbe, X.; Venu, B.; Jagst, J.; Ozer, E.; Harrod, P.; Turner, C.; Penton, J. Addressing Functional Safety Challenges in Autonomous Vehicles with the Arm TCL S Architecture. *IEEE Des. Test* **2018**, *35*, 7–14. [CrossRef]
8. Patel, A.R.; Liggesmeyer, P. Machine Learning Based Dynamic Risk Assessment for Autonomous Vehicles. In Proceedings of the 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Rome, Italy, 12–14 November 2021; pp. 73–77. [CrossRef]
9. Li, G.; Yang, Y.; Zhang, T.; Qu, X.; Cao, D.; Cheng, B.; Li, K. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. *Transp. Res. Part C Emerg. Technol.* **2021**, *122*, 102820. [CrossRef]
10. Drage, T.; Lim, K.L.; Koh, J.E.H.; Gregory, D.; Brogle, C.; Braunl, T. Integrated modular safety system design for intelligent autonomous vehicles. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 258–265. [CrossRef]
11. Khalil, A.; Al Janaideh, M.; Aljanaideh, K.F.; Kundur, D. Fault Detection, Localization, and Mitigation of a Network of Connected Autonomous Vehicles Using Transmissibility Identification. In Proceedings of the 2020 American Control Conference (ACC), Denver, CO, USA, 1–3 July 2020; pp. 386–391. [CrossRef]
12. Shetty, A.; Tavafoghi, H.; Kurzhanskiy, A.; Poolla, K.; Varaiya, P. Risk Assessment of Autonomous Vehicles across Diverse Driving Contexts. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 712–719. [CrossRef]
13. Shen, Y.; Xia, C.; Jian, Z.; Chen, S.; Zheng, N. An Integrated Localization System with Fault Detection, Isolation and Recovery for Autonomous Vehicles. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 84–91. [CrossRef]
14. Jacumet, R.; Rathgeber, C.; Nenchev, V. Analytical Safety Bounds for Trajectory Following Controllers in Autonomous Vehicles. In Proceedings of the 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT), Rome, Italy, 3–6 July 2023; pp. 730–735. [CrossRef]
15. Hashemi, M.; Golkani, M.A.; Watzenig, D. A Robust Approach for Inter-Turn Fault Detection of PMSM Used for Autonomous Vehicles. In Proceedings of the 2022 International Conference on Connected Vehicle and Expo (ICCVE), Lakeland, FL, USA, 7–9 March 2022; pp. 1–6. [CrossRef]
16. Sheikh, M.S.; Peng, Y. Improved Collision Risk Assessment for Autonomous Vehicles at on-Ramp Merging Areas. *IEEE Access* **2023**, *11*, 130974–130989. [CrossRef]

17. Liu, Z.; Liu, X.; Li, Q.; Zhang, Z.; Gao, C.; Tang, F. Strategies for Coordinated Merging of Vehicles at Ramps in New Hybrid Traffic Environments. *Sustainability* **2025**, *17*, 4522. [CrossRef]
18. Liang, J.; Yang, K.; Tan, C.; Wang, J.; Yin, G. Enhancing high-speed cruising performance of autonomous vehicles through integrated deep reinforcement learning framework. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 835–848. [CrossRef]
19. Jeong, Y. Fault detection with confidence level evaluation for perception module of autonomous vehicles based on long short term memory and Gaussian Mixture Model. *Appl. Soft Comput.* **2023**, *149*, 111010. [CrossRef]
20. Madala, K.; Do, H. Functional safety hazards for machine learning components in autonomous vehicles. In Proceedings of the 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS), Victoria, BC, Canada, 10–12 May 2021; pp. 225–230. [CrossRef]
21. Tan, H.; Zhao, F.; Zhang, W.; Liu, Z. An Evaluation of the Safety Effectiveness and Cost of Autonomous Vehicles Based on Multivariable Coupling. *Sensors* **2023**, *23*, 1321. [CrossRef] [PubMed]
22. Bernhard, J.; Hart, P.; Sahu, A.; Scholler, C.; Cancimance, M.G. Risk-Based Safety Envelopes for Autonomous Vehicles Under Perception Uncertainty. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; pp. 104–111. [CrossRef]
23. Stefanoni, M.; Takács, M.; Odry, Á.; Sarcevic, P. A Comparison of Neural Networks and Fuzzy Inference Systems for the Identification of Magnetic Disturbances in Mobile Robot Localization. *Acta Polytech. Hung.* **2025**, *22*, 239–264. [CrossRef]
24. Zhang, S.; Puig, V.; Ifqir, S. Robust Fault Detection using Set-based Approaches for LPV Systems: Application to Autonomous Vehicles. *IFAC-PapersOnLine* **2022**, *55*, 31–36. [CrossRef]
25. Khalil, A.; Al Janaideh, M.; Aljanaideh, K.F.; Kundur, D. Output-only fault detection and mitigation of networks of autonomous vehicles. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 2257–2264. [CrossRef]
26. Fu, Y.; Terechko, A.; Bijlsma, T.; Cuijners, P.J.L.; Redegeld, J.; Ors, A.O. A Retargetable Fault Injection Framework for Safety Validation of Autonomous Vehicles. In Proceedings of the 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), Hamburg, Germany, 25–26 March 2019; pp. 69–76. [CrossRef]
27. Leng, B.; Yu, R.; Han, W.; Xiong, L.; Li, Z.; Huang, H. Risk-Aware Reinforcement Learning for Autonomous Driving: Improving Safety When Driving through Intersection. *arXiv* **2025**. [CrossRef]
28. Liang, J.; Tian, Q.; Feng, J.; Pi, D.; Yin, G. A polytopic model-based robust predictive control scheme for path tracking of autonomous vehicles. *IEEE Trans. Intell. Veh.* **2023**, *9*, 3928–3939. [CrossRef]
29. Yang, S.; Du, M.; Chen, Q. Impact of connected and autonomous vehicles on traffic efficiency and safety of an on-ramp. *Simul. Model. Pract. Theory* **2021**, *113*, 102374. [CrossRef]
30. Wang, J.; Jin, Y.; Taghavifar, H.; Ding, F.; Wei, C. Socially-Aware Autonomous Driving: Inferring Yielding Intentions for Safer Interactions. *arXiv* **2025**. [CrossRef]
31. Lyu, Y.; Luo, W.; Dolan, J.M. Probabilistic Safety-Assured Adaptive Merging Control for Autonomous Vehicles. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 10764–10770. [CrossRef]
32. Wang, K.; Shen, C.; Li, X.; Lu, J. Uncertainty Quantification for Safe and Reliable Autonomous Vehicles: A Review of Methods and Applications. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 2880–2896. [CrossRef]
33. Tan, H.; Zhao, F.; Song, H.; Liu, Z. Quantifying the Impact of Deployments of Autonomous Vehicles and Intelligent Roads on Road Safety in China: A Country-Level Modeling Study. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4069. [CrossRef] [PubMed]
34. Shao, Y.; Han, Z.; Shi, X.; Zhang, Y.; Ye, Z. Risk-informed longitudinal control in autonomous vehicles: A safety potential field modeling approach. *Phys. A Stat. Mech. Its Appl.* **2024**, *633*, 129419. [CrossRef]
35. *ISO 26262:2018*; Road Vehicles—Functional safety (Parts 1–12). International Organization for Standardization: Geneva, Switzerland, 2018.
36. *ISO/PAS 21448:2019*; Road Vehicles—Safety of the Intended Functionality (SOTIF). International Organization for Standardization: Geneva, Switzerland, 2019.
37. Saulaiman, M.N.E.; Csilling, A.; Kozlovszky, M. Integrated Automation for Threat Analysis and Risk Assessment in Automotive Cybersecurity Through Attack Graphs. *Acta Polytech. Hung.* **2025**, *22*, 149–168. [CrossRef]
38. *ISO/SAE 21434:2021*; Road Vehicles—Cybersecurity Engineering. International Organization for Standardization & SAE International: Geneva, Switzerland, 2021.
39. Yu, R.; Wang, C.; Zhang, Y.; Zhao, F. Decomposition and Quantification of SOTIF Requirements for Perception Systems of Autonomous Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2025**, *early access*. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.