OPEN ACCESS
IOP Publishing | ISECAEP
Journal of Reliability Science and Engineering
J. Reliab. Sci. Eng. **2** (2026) 012002 (27pp)
https://doi.org/10.1088/3050-2454/ae4d4b

Topical Review

Review on reliability of intelligent autonomous systems

Jie Liu[1], Shuwen Zheng[1,2], Yunxia Chen[1,2,*], Dan Xu[1], Cong Wang[1,2], Weiyi Xiang[1,2], Xiaoqi Xiao[1] and Jing Lin[1,2,*]

[1] School of Reliability and Systems Engineering, Beihang University, Beijing 100191, People's Republic of China
[2] State Key Laboratory of Advanced Forming Technology and Equipment, Beihang University, Beijing 100191, People's Republic of China

E-mail: chenyunxia@buaa.edu.cn and linjing@buaa.edu.cn

Received 16 December 2024, revised 31 January 2026
Accepted for publication 3 March 2026
Published 27 March 2026

Abstract

Intelligent autonomous systems (IASs), encompassing autonomous vehicles, unmanned aerial vehicles, and robotic platforms, have revolutionized sectors ranging from transportation to defense. By leveraging artificial intelligence (AI) techniques, these systems are characterized by data-driven learning paradigms, integrated architectures, and adaptive decision-making. However, these novel capabilities introduce distinctive failure challenges, including data-induced biases, limited interpretability, and vulnerability to adversarial perturbations. Such features necessitate enhanced reliability to ensure dependable operation in dynamic, safety-critical environments. This review synthesizes the state-of-the-art in IAS reliability engineering, framed through four interconnected dimensions: accuracy, generalization, robustness, and explainability. Drawing from interdisciplinary advancements in machine learning, AI and reliability, we examine methodologies such as multi-sensor fusion, meta-learning, adversarial training, and ante-hoc interpretability constraints, supported by empirical evidence from real-world deployments. Key insights highlight substantial progress in mitigating IASs vulnerabilities, yet persistent challenges including performance-reliability trade-offs, degradation under extreme conditions, and scalability limitations impede widespread adoption. We propose future directions emphasizing hybrid frameworks, causal inference and lightweight models to advance reliable IASs. By bridging theoretical foundations with practical implementations, this work provides a comprehensive roadmap for developing reliable and trustworthy autonomous systems that prioritize safety, efficiency, and societal well-being.

Keywords: autonomous systems, reliability engineering, artificial intelligence

* Authors to whom any correspondence should be addressed.

1. Introduction

Driven by rapid technological advancements, intelligent autonomous systems (IASs), such as autonomous unmanned aerial vehicles (UAVs), self-driving cars, and autonomous ships, have become increasingly integrated into modern society, revolutionizing sectors like transportation, logistics, agriculture, and defense [1, 2]. Notable applications include Waymo's autonomous taxis in the United States, which had accumulated over 71 million miles of autonomous driving by 2025 [3]. In China, Baidu's Apollo platform has enabled autonomous taxi services, completing over 11 million rides across 15 Chinese cities [4]. In the maritime domain, Norway's Yara Birkeland, the world's first fully electric and autonomous container ship, began commercial operations recently, showcasing significant efficiency gains in short-sea shipping. These systems are characterized by their ability to perceive the environment, make decisions, and execute actions independently, leveraging diverse decision-making approaches and advanced machine learning (ML) models [5].

Typically, IASs incorporate perception modules (e.g. light detection and ranging (LiDAR), cameras, radar), decision-making algorithms (e.g. path planning, optimization), and execution/control components (e.g. actuators, propulsion systems), enabling them to achieve predefined objectives while dynamically adapting to changing conditions [6]. Autonomous systems can be broadly classified into rule-based and learning-based systems, with the latter prevailing in modern IASs. These learning-based systems harness sophisticated artificial intelligence (AI) models, such as convolutional neural networks (CNNs) for image processing, long short-term memory (LSTM) networks for sequential data analysis, attention mechanisms for contextual awareness, and reinforcement learning (RL) algorithms for adaptive decision-making, thereby enhancing performance in dynamic and uncertain environments [7–9]. As these sophisticated AI algorithms are progressively embedded, IASs demonstrate **novel technological features** that distinguish them from traditional engineered systems. First, they leverage ***data-driven learning paradigms***, allowing models to extract intricate patterns from vast datasets that would be infeasible for human or traditional analysis, enabling superior handling of complex scenarios. Second, IASs feature ***integrated end-to-end architectures*** that fuse perception, reasoning, and action into seamless processes, promoting high efficiency in multifaceted environments without the constraints of predefined rules. Third, they demonstrate ***autonomous task reconfigurability and environmental adaptability***. This allows IASs to not only adjust strategies in real time but also switch between operational modes or objectives based on dynamic feedback, achieving unprecedented levels of autonomy and resilience in rapidly changing settings. Collectively, these models and features position IASs as pivotal drivers of automation, efficiency, and innovation across diverse applications.

Nevertheless, the widespread adoption of IASs underscores the urgent need to prioritize their reliability. Often deployed in safety-critical settings, these systems are exposed to unforeseen uncertainties that can trigger catastrophic failures, ranging from operational breakdowns to threats to human life [10, 11]. This risk is compounded by the inherent automation of IASs, which fosters greater user dependency and reduces opportunities for timely manual intervention or error correction [12]. In real-world applications, these concerns manifest through **distinct failure characteristics** unique to IASs. First, their ***heavy reliance on data and data-driven models*** such as deep learning and RL, makes them vulnerable to biases arising from poor data quality, or distribution shifts. These issues often lead to ***compromised generalization and transferability***, exacerbated by factors like data imbalance, over/underfitting, or suboptimal hyperparameter selection, impeding adaptation to novel environments or related tasks and, ultimately, compromising overall system performance and reliability [13]. Additionally, IASs often exhibit ***poor model interpretability***, where intricate nonlinear transformations render decision-making processes opaque, akin to 'black boxes', significantly impairing trust, accountability, and debugging in high-stakes scenarios [14]. These challenges are further intensified by ***susceptibility to adversarial attacks***, which can subtly manipulate sensor inputs to provoke erroneous judgments and inadequate responses [15]. Real-world incidents have starkly highlighted the associated risks: a fatal accident involving an Uber self-driving car in Arizona stemmed from a sensor data misclassification that failed to detect a pedestrian [16], while a Boeing A160T Hummingbird prototype crashed during a flight test due to lost sensor feedback, which disrupted the stabilization control loop and led to a near-vertical impact [17]. Such inherent challenges and events highlight the imperative to advance reliability engineering, ensuring IASs operate dependably across diverse and unpredictable real-world conditions.

Formally, reliability can be defined as the ability of a system to consistently perform its intended functions under specified conditions without failure [18]. The growing application of IASs has driven extensive research to enhance the reliability of autonomous and AI-driven systems, encompassing aspects such as reliability analysis, risk management, and uncertainty quantification (UQ) [19]. Several review papers have focused on these challenges from diverse angles. For example, Yu *et al* [20] presented a taxonomy of AI system failures and surveyed failure analysis and fault injection methods, evaluating existing tools and underscoring gaps between real-world failures and simulated scenarios to advocate for robust testing frameworks. Osborne *et al* [21] focused on reliability challenges in unmanned aerial systems (UASs), proposing solutions such as integrated vehicle health management, simulations and fault tolerant control to improve operational reliability. Similarly, Blood *et al* [22] examined reliability assurance, noting that traditional methods, such as failure modes and effects analysis (FMEAs) and reliability block diagrams, remain vital but require adaptation for AI-specific risks. In addition, Olamide *et al* [23] emphasized the pivotal role of component integrity, recommending proactive strategies like probabilistic fault detection models to minimize failures and improve reliability of autonomous systems. Finally, Flammini *et al* [24] mapped

key concepts in trustworthy AI, identifying emerging challenges and calling for structured frameworks and regulatory progress.

While these prior reviews have provided valuable insights, they predominantly focus on system-level hardware reliability or specific testing methodologies [20, 21, 23]. However, the reliability of modern IASs is increasingly governed by the logical correctness and adaptability of their intelligent algorithms. Therefore, this review differs in scope by explicitly focusing on AI-driven decision-making reliability rather than system-level hardware reliability. In this paper, we present an in-depth review grounded in the novel technological and failure characteristics inherent to these systems. We emphasize that the integration of advanced AI models, characterized by features such as data-driven learning paradigms, integrated architectures and adaptability, introduces unique reliability demands. Distinctive failure modes including perception errors stemming from data bias, decision instability under adversarial attacks and risks exacerbated by poor interpretability and limited generalization, constitute threats that cannot be addressed by system-level hardware redundancy alone. Consequently, this work focuses on the decision-making processes of intelligent algorithms and systems, seeking to provide a cohesive overview of IAS reliability framed through four interconnected algorithmic dimensions: accuracy (e.g. precise perception and decision-making), generalization (adaptation across unseen scenarios), robustness (resilience to uncertainties and attacks), and explainability/transparency (interpretable decision logic). By synthesizing underlying principles, state-of-the-art methodologies, and empirical insights from diverse applications, we aim to deliver valuable perspectives that bridge theoretical foundations with practical advancements, ultimately fostering the development of more reliable, trustworthy, and resilient IASs.

The remainder of the paper is structured as follows. Section 2 defines IASs reliability in detail and offers an in-depth examination of current technologies and research in reliability engineering, covering accuracy, generalization, robustness, and explainability. Section 3 discusses key challenges such as the intrinsic trade-off and computational complexity, and proposes promising research directions with supporting case studies. Finally, section 4 concludes the paper with a synthesis of insights and recommendations for future work.

2. Current reliability engineering of IASs

2.1. Definition

Reliability in IASs refers to their ability to consistently perform intended tasks without failure, even when subjected to dynamic and unpredictable environments. Unlike traditional systems, which often rely on deterministic models and pre-programmed responses, IASs leverage complex ML and DL techniques to adapt to and learn from their surroundings. As IASs increasingly depend on advanced AI technologies, their reliability differs significantly from that of conventional systems. In traditional reliability engineering, time-dependence is often characterized by physical wear-out, e.g. modeled via the bathtub curve [18]. Conversely, the reliability of IASs is typically governed by the evolution of the operational environment. For example, the statistical distribution of real-world data may drift from the training environment over time, causing the reliability to degrade continuously. Furthermore, in continuous operation, errors in decision-making can accumulate, leading to mission failure over extended periods.

Therefore, IAS reliability must be viewed as a dynamic property that requires sustained adaptation, accounting for the complexity, adaptability, and unique failure modes of AI models. These include risks such as data biases, model overfitting, and vulnerability to adversarial manipulation, all of which pose threats to system reliability. Achieving this requires a thorough understanding of the interaction between AI techniques and system performance, as well as the development of strategies to mitigate the specific risks introduced by the integration of AI into autonomous systems. Collectively, the reliability of IASs includes several key dimensions:

- **Accuracy:** The extent to which system outputs match expected results in given context. In the context of AI, accuracy refers to the system's ability to make precise predictions and decisions based on sensory input and environmental data.
- **Generalization:** Since real-world conditions often diverge from training environments, IASs must extend beyond the data and scenarios they were originally trained on. Generalization reflects the system's capacity to adapt learned models to new tasks, environments, or unforeseen disturbances, thereby ensuring consistent performance across diverse situations.
- **Robustness:** AI models can be vulnerable to perturbations in input data or environmental factors. Robustness denotes the system's ability to sustain functional performance despite exposure to uncertainties, noise, adversarial attacks, and environmental variations.
- **Explainability/transparency:** The decision-making processes of AI models in IASs can often be opaque, making it difficult for users to understand why a system made a particular decision. Reliability is thus closely linked to explainability, which fosters trust, accountability, and the ability to diagnose errors or failures.

Together, these dimensions collectively define the reliability of IASs, ensuring not only functional correctness but also trust and safety in real-world operations. As IASs evolve, addressing the challenges inherent to each dimension remains essential for enhancing their reliability in complex, high-stakes environments. To facilitate a holistic understanding, figure 1 presents a conceptual taxonomy mapping that visually summarizes key enhancement strategies across the four interconnected dimensions, categorizing methodologies into primary approaches with representative examples and citations. This framework guides the subsequent detailed reviews in sections 2.2–2.5. Arrows in the mapping highlight interdependencies and potential trade-offs, such as shared uncertainty handling between generalization and robustness, and the robustness-accuracy dilemma, which are further elaborated in

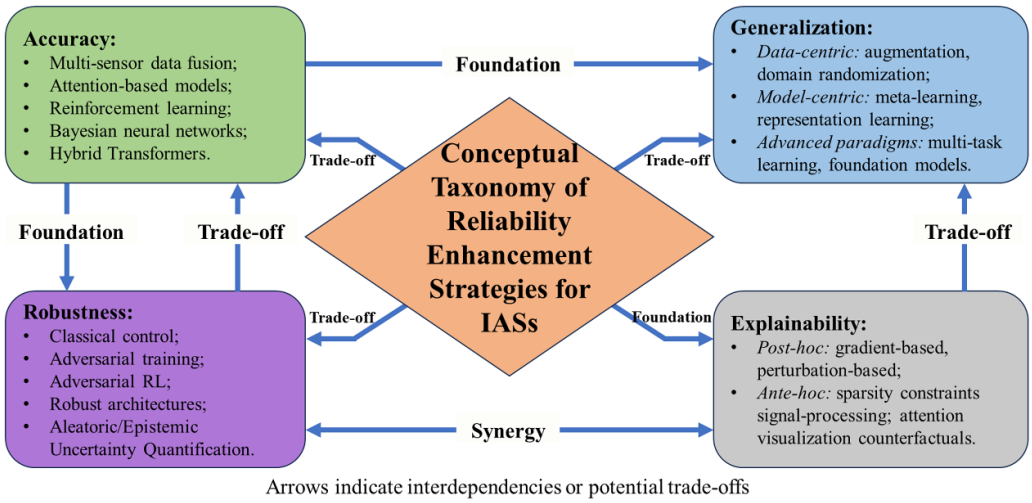


Figure 1. Conceptual taxonomy mapping of reliability enhancement strategies in IASs across dimensions.

the section 2.6, enabling readers to contextualize state-of-the-art approaches within IAS reliability engineering.

2.2. Accuracy

2.2.1. Multidimensional accuracy: perception, prediction and control. The accuracy of IASs refers to the consistency between expected and actual outputs in given task scenarios, serving as a fundamental indicator of system reliability. As IASs are deployed across diverse domains, accuracy has become a multifaceted concept spanning perception, prediction, and control. In perception tasks such as multiple object tracking, the multiple object tracking accuracy (MOTA) is widely adopted, defined as

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \mathrm{IDSW}}{\mathrm{GT}} \tag{1}$$

where FN, FP, IDSW, and GT represent the numbers of false negatives, false positives, identity switches, and ground truths, respectively [25]. Improving MOTA can lower the risk of downstream planning failures due to erroneous environmental understanding, thereby enhancing operational reliability. In prediction, metrics such as the minimum final displacement error (minFDE) evaluate long-term trajectory forecasting; among a set of K predicted future trajectories for a target, minFDE is the Euclidean distance between the endpoint of the ground truth trajectory and the endpoint of the closest predicted trajectory:

$$\mathrm{minFD}\mathbb{E}_K = \min_{k=1,\cdots,K} \left\| y_T - \hat{y}_T^{(k)} \right\|_2 \tag{2}$$

where y_T and $\hat{y}_T^{(k)}$ denote the final coordinates of the ground truth and the kth predicted trajectory, respectively, and $\|\cdot\|_2$ is the L2-norm [26]. A reduction in minFDE, for example, can decrease the probability of collisions caused by positional estimation errors, which can improve reliability metrics such as the mean time between failures (MTBFs). In control, accuracy is often assessed by position drift (the Euclidean distance between the actual and target final position) and rotation deviation (the minimal angular difference between the actual and target orientation) [27]; minimizing these deviations is essential for ensuring consistent operational performance and long-term system stability. Thus, accuracy metrics serve as quantifiable proxies for reliability attributes of IASs.

2.2.2. Accuracy enhancement paradigms. To enhance the accuracy of IASs, researchers have developed diverse approaches spanning critical technical areas such as perception, prediction, and control. This section offers a detailed review of representative studies, highlighting their core innovations, application contexts, strengths, and limitations. Table 1 summarizes these studies for accuracy improvement.

Accurate perception serves as the foundation for reliable prediction and control in IASs. Achieving it in engineering necessitates trade-offs between performance and efficiency, depending on application-specific priorities. In safety-critical domains like autonomous driving, accuracy under harsh operating conditions is paramount, which often leads to the adoption of computationally intensive but highly accurate models. As illustrated by Walambe *et al* [28], an integrated object detection framework that leveraged the rapid response of YOLOv3 and the small-object sensitivity of RetinaNet achieved significant accuracy improvement over baseline models under adverse weather conditions; this integrated approach itself provides a form of technical redundancy against environmental uncertainty. However, such an approach incurs substantial processing overhead, making it less suitable for severely resource-constrained platforms. Conversely, for persistent monitoring applications where long-term deployability and energy efficiency take precedence, researchers often focus on architectural light-weighting. For example, Musabimana *et al* [29] developed an enhanced lightweight

Table 1. Representative studies for improving accuracy in IASs.

Category	Representative works	Key contributions	Application domain
Perception	Walambe *et al* [28]	YOLOv3 + RetinaNet, better detection in bad weather	Autonomous driving
	Musabimana *et al* [29]	Lightweight hybrid Transformer, 5.99% RMSE reduction	Smart agriculture
Prediction	Cheng *et al* [30]	Graph + GCN for pedestrian trajectory	Autonomous driving
	Lu *et al* [31]	Transformer transfer learning for lane-change	Autonomous driving
	Zou *et al* [32]	Unified model with pose/trajectory/group context	Autonomous driving
Control	Ma *et al* [33]	Deep RL for accurate attitude/velocity control	UAVs
	Xia *et al* [34]	Soft actor-critic, <0.1 m lateral error	Autonomous driving
	Ran *et al* [35]	Driving-style car-following, 94.8% success	Autonomous driving
	Nan *et al* [36]	Inverse RL with attention, collision risk 0.01%	Autonomous driving
	Lebede and Nadarajah [37]	Bayesian NN for uncertainty, errors reduced 32%–42%	Autonomous driving
	Yang *et al* [38]	RL control for blast stoves, 85.9% accuracy	Smart manufacturing
	Liu [39]	Vision-based polishing with Mask R-CNN, <1 mm error	Smart manufacturing

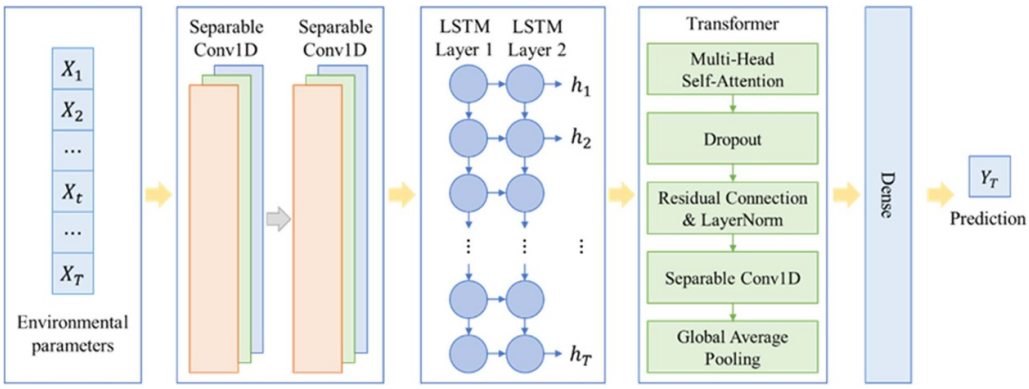


Figure 2. Structure of the lightweight hybrid transformer model for intelligent monitoring.

hybrid Transformer model incorporating depthwise separable convolutions and global average pooling for intelligent environmental monitoring (figure 2). It achieves dramatic gains in operational efficiency, reducing model size by 21.1% and training time by 29.7% while maintaining competitive prediction accuracy. Thus, the pursuit of accurate perception entails a balance: leveraging computational resources for maximal reliability in dynamic environments, or prioritizing sustainable efficiency for stable, long-term deployments.

Accurate prediction is another cornerstone for reliable decision-making and control in IASs, with trajectory prediction emerging as one of the most extensively studied challenges. Current research varies in its approach to handling the inherent complexities and uncertainties of prediction, particularly in terms of data utilization, scene dependency, and interaction modeling. Cheng *et al* [30] proposed an attention-based graph model for pedestrian trajectory prediction. Its scene-centric graph design efficiently captures spatiotemporal interactions in open environments, achieving a favorable accuracy-speed balance. However, its performance is constrained in highly structured environments, as it intentionally omits scene semantics such as lane boundaries, which can limit the ability to manage environmental uncertainty and lead to physically implausible predictions (figure 3). In contrast, Lu *et al* [31] addressed the data scarcity issue in lane-change scenarios by introducing a transfer learning framework. This method leverages large-scale, generic trajectory data to pre-train a model, creating a form of knowledge redundancy by learning complex inter-vehicle interaction patterns, which are then transferred to boost the performance of a dedicated lane-change prediction model. While this significantly improves accuracy against data sparsity, it adds complexity to the training phase. Zou *et al* [32] tackled complexity in pedestrian intention prediction by fusing multiple data modalities including pose, trajectory, and social group context. Their work demonstrates that explicitly modeling group behavior yields significant accuracy gains in complex scenarios. However, this approach introduces a trade-off: it creates a strong dependency on the robustness of upstream feature extractors, such as pose estimation models, making the system vulnerable to failures under occlusion or poor sensing conditions. Together, these studies highlight how advanced graph-based, Transformer-based, and multi-modal learning architectures are advancing the predictive capabilities of IASs across diverse dynamic environments.

Control accuracy is a critical determinant of IAS reliability. Research on improving control accuracy in IASs reveals distinct technical pathways shaped by application scenarios and

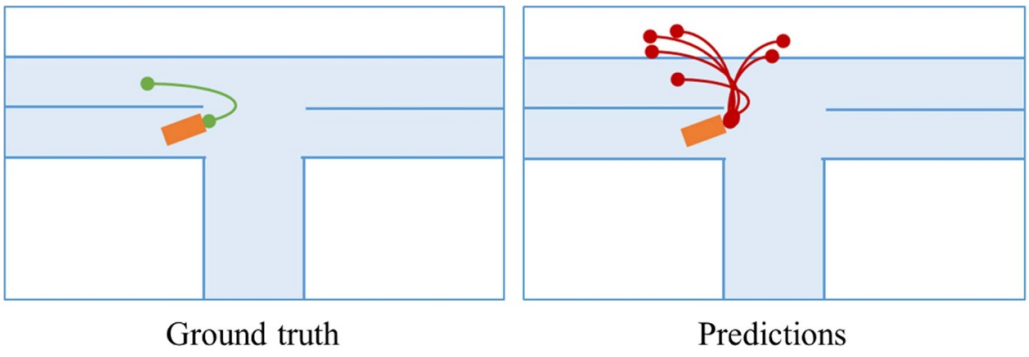


Figure 3. Conceptual illustration of physically implausible trajectory predictions when omitting scene semantics.

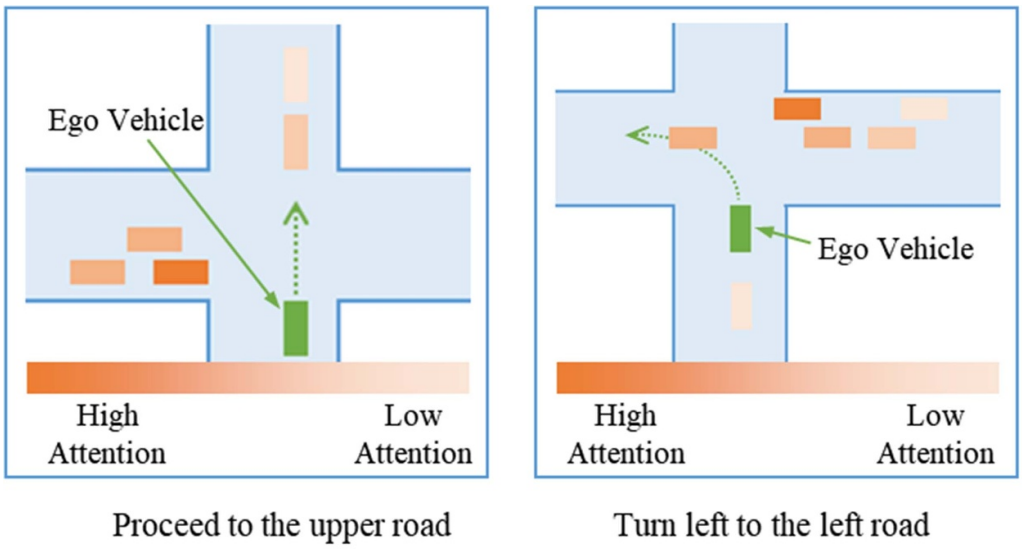


Figure 4. Conceptual illustration of human-like attention allocation of road vehicles.

core challenges. For handling dynamic uncertainties, Deep RL (DRL) based controllers demonstrate strong adaptability. For instance, Ma *et al* [33] and Xia *et al* [34] designed DRL controllers for UAV wind disturbance rejection and high-curvature vehicle path tracking, respectively, achieving high precision while validating from simulation to hardware-in-the-loop. However, such methods generally require extensive training data and intricate reward shaping, and their policies lack interpretability. To enhance decision-making trust and human-likeness, more complex modeling mechanisms are introduced. Nan *et al* [36] utilized inverse RL with attention mechanisms to mimic human driver preferences (figure 4), significantly improving behavioral realism and safety, but their model training is more complex and heavily reliant on the quality of demonstration data. Similarly, Lebede and Nadarajah [37] employed Bayesian neural networks (BNN) to explicitly quantify decision uncertainty, enabling more conservative and safer driving in adverse conditions, albeit at a high computational cost. In structured industrial settings, hybrid approaches combining learning with traditional control are prevalent. Yang *et al* [38] applied RL to optimize blast furnace valve control, addressing processes difficult to model analytically, while Liu [39] integrated Mask R-CNN vision with PID-impedance control for robotic polishing, achieving sub-millimeter accuracy. These hybrid schemes often incorporate algorithmic redundancy, merging data-driven adaptation with deterministic control laws to enhance overall system reliability, but they often lack generalization beyond specific operational environments.

While all the above strategies aim to enhance accuracy, their implications for reliability assurance vary. Purely deep learning-based perception enhancements, such as the integrated object detection framework, provide high performance but introduce 'black-box' non-determinism, making safety certification (e.g. ISO 26262/SOTIF) challenging in safety-critical scenarios [28]. In contrast, hybrid control strategies that combine vision with deterministic PID or impedance control offer superior verifiability, as the control loop remains analytically stable even with component degradations. Furthermore, while lightweight architectures reduce latency, they often lack the structural redundancy of multi-modal fusion models. Given these trade-offs, relying solely on deterministic accuracy metrics is increasingly insufficient. Consequently, for high-stakes applications like autonomous driving, there is a clear trend towards methods that not only maximize accuracy metrics, e.g. MOTA, but also incorporate

Table 2. Domain generalization strategies for IASs: advantages and limitations.

Strategies	Approach	Advantages	Limitations
Data-centric strategies	Data augmentation	Simple, low-cost	May be ineffective or introduce bias
	Domain randomization	Simple and effective	Hard to design; may not match real-world
Model-centric strategies	Meta-learning	Rapid adaptation with few samples	Needs diverse tasks; computationally heavy
	Representation learning	Learns transferable robust features	Depends on source–target similarity
Advanced learning paradigms	Multi-task learning	Improves cross-task generalization	Unrelated tasks may hurt performance
	Large-scale pre-trained models	Strong zero-/few-shot generalization	Requires huge compute/data; low interpretability

explicit uncertainty bounds, serving as a necessary condition for reliability assessment [40].

2.3. Generalization

IASs often face significant challenges of performance consistency when transitioning from controlled training environments to the complex and dynamic open world [41]. A core issue lies in the system's generalization capability. Generalization refers to the ability of autonomous intelligent agents to transfer and apply existing knowledge and skills to unknown environments and tasks, which is crucial for ensuring the reliability and safety of IASs in the open world [42].

2.3.1. Domain divergence and theoretical risk bounds. In AI, generalization denotes a model's ability to sustain performance on unseen data or tasks. For IASs, this manifests as the capacity to make accurate decisions and execute tasks under varying environmental conditions, sensory inputs, or task requirements. For instance, an autonomous vehicle trained on daytime data in clear weather must demonstrate generalization by safely navigating out-of-distribution (OOD) scenarios, such as nighttime or foggy conditions. Strong generalization enables IASs to adapt existing knowledge to diverse scenarios, showcasing environmental adaptability and task transferability.

Given M training (source) domains $S_{\text{train}} = \{S^i | i = 1, \cdots, M\}$, where $S^i = \left\{\left(\mathbf{x}_j^i, y_j^i\right)\right\}_{j=1}^{n_i}$ represents the ith domain. The joint distribution between domains differs: $P_{XY}^i \neq P_{XY}^j$, $1 \leqslant i \neq j \leqslant M$. The goal of domain generalization can be formalized as learning a robust and generalizable prediction function $h : \mathcal{X} \to \mathcal{Y}$ from the training domains that minimizes prediction error on an unseen test domain S_{test} [43]:

$$\min_h \mathbb{E}_{(x,y)\in S_{\text{test}}} \left[\ell\left(h\left(x\right), y\right)\right] \tag{3}$$

where $\mathbb{E}$ represents expectation, $\ell\left(\cdot,\cdot\right)$ is the loss function.

The expected risk on a target domain $\varepsilon_{\text{T}}\left(h\right)$ is bounded by the source risk $\varepsilon_{\text{S}}\left(h\right)$ and the domain divergence:

$$\varepsilon_{\text{T}}\left(h\right) \leqslant \varepsilon_{\text{S}}\left(h\right) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}\left(D_{\text{S}}, D_{\text{T}}\right) + \lambda \tag{4}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the discrepancy between distributions D, and λ represents the ideal joint risk. This theoretical framework reveals that the core challenge of DG is minimizing the divergence $d_{\mathcal{H}\Delta\mathcal{H}}$ without compromising the discriminability required to keep $\varepsilon_{\text{S}}\left(h\right)$ low.

2.3.2 Strategies for cross-domain generalization. To address the challenges of generalization, researchers have proposed a range of solutions from various dimensions, such as data, models, and learning paradigms. This section categorizes these methods into three main strategies and provides a detailed discussion. Table 2 summarizes these strategies for IASs and contrasts their key strengths and limitations.

A. Data-centric strategies

A straightforward method to improve generalization is to enhance the diversity and coverage of training data, exposing models to varied scenarios. Key techniques include data augmentation and domain randomization.

Data augmentation generates synthetic training samples through transformations of existing data, expanding dataset diversity without additional collection. Basic augmentation, common in computer vision, includes geometric transformations (e.g. random rotations, scaling, cropping) and color space adjustments (e.g. brightness, contrast, saturation) [44]. These operations simulate the appearance of objects under different angles, distances, and lighting conditions. Advanced augmentation aims to create semantically challenging samples. For example, generative augmentation methods use generative adversarial networks [45] or variational autoencoders [46] to learn the latent distribution of the training data and generate new, highly realistic synthetic data. Another method, Mixup [47], generates new samples by linearly interpolating between two randomly selected samples and their labels as:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{5}$$

where (x_i, y_i) and (x_j, y_j) are two training samples, and λ is a mixing coefficient sampled from a Beta distribution. Data augmentation is simple, cost-effective, and effective for limited datasets, but basic transformations may not produce novel semantic samples (e.g. rotating a car image cannot yield a

truck). Poorly designed augmentations may also introduce biases misaligned with real-world data, potentially degrading performance.

While data augmentation is the most cost-effective method for addressing standard visual corruptions (e.g. lighting, noise) in limited datasets, it is often insufficient for complex physical interactions. Under conditions involving sim-to-real transfer or dynamics mismatches, **domain randomization** is significantly more effective. By randomizing physical parameters, it prevents the policy from overfitting to the simulation. Therefore, domain randomization is preferable when the primary domain shift arises from environmental physics rather than mere visual appearance. The essence of domain randomization is to produce diverse training samples by randomly perturbing environmental parameters such as material textures and physical parameters in simulated environments [48]. This approach ensures that the trained policy does not overfit to a specific environment but instead adapts to a variety of unstructured conditions, making it more capable of handling new situations. For example, OpenAI's robotic grasping tasks randomized object colors, backgrounds, and lighting, allowing policies to generalize to real-world settings. Chen *et al* [49] proposed adversarial domain randomization, using an adversarial generator to create realistic training environments for complex scenarios. While simple and effective, especially in RL, domain randomization requires careful design of perturbation ranges. Inadequate ranges may fail to cover real-world variations, while excessive randomization can complicate learning. It also struggles with unconsidered real-world factors, such as novel sensor interference [50].

From a reliability engineering perspective, domain randomization functions as a simulation-based reliability assessment framework. By systematically introducing environmental 'faults' and sensor disturbances during training, it effectively acts as a proactive stress test, identifying potential failure modes within the operational design domain (ODD) before real-world deployment.

The above data-centric strategies serve as a low-cost baseline effective for sensor-level noise robustness. However, they face a 'semantic gap' limitation: basic transformations cannot simulate high-level semantic shifts (e.g. changing a car to a truck), and aggressive randomization may introduce unrealistic biases that degrade real-world performance. These methods are most applicable to closed-loop control tasks (e.g. robotic grasping) where physical parameters are the primary variable, but struggle with complex open-world semantic changes.

B. Model-centric strategies

These strategies focus on improving the model architecture or learning objectives themselves, endowing the model with stronger generalization and adaptability. Key approaches include meta-learning and representation learning.

Meta-learning, often referred to as 'learning how to learn', aims to enable models to adapt quickly to new tasks with minimal data. For IASs, it targets task-level generalization, serving as a key enabler for autonomous task reconfigurability. By allowing agents to master new tasks explicitly with minimal data, meta-learning ensures reliability when the system must reconfigure its objectives in unforeseen scenarios. A classic meta-learning algorithm is model-agnostic meta-learning (MAML) [51], which seeks to optimize initial model parameters θ for high 'plasticity', achieving strong performance with few gradient updates on new tasks. The optimization objective can be defined as:

$$\min_{\theta} \sum_{T_i \sim p(T)} L_{T_i}\left(f_{\theta_i'}\right) \text{ where } \theta_i' = \theta - \alpha \nabla_{\theta} L_{T_i}\left(f_{\theta}\right) \tag{6}$$

where $p(T)$ represents the distribution of tasks. The algorithm optimizes the meta-parameters θ in the outer loop and simulates the rapid adaptation process to a new task T_i in the inner loop.

In addition to MAML, many variants have been proposed for the fields of robotics and RL. Yu *et al* [52] established a meta-learning benchmark based on 50 robotic arm manipulation tasks, promoting standardized evaluation of algorithms. Bao *et al* [53] introduced the memory augmentation strategy, which applies structured task perturbations to the experience during meta-training, simulating potential OOD tasks. By utilizing recurrent neural networks (RNNs) to implicitly infer this latent task context from history, this approach enables zero-shot generalization to new tasks. Figure 5 shows an overview of the proposed framework. They demonstrated this in a robot walking task, where the robot was able to directly adapt to a more complex new situation without additional training, demonstrating robust adaptability. This capability essentially provides a mechanism for software-defined adaptive fault tolerance. Unlike traditional hardware redundancy, the meta-learning agent implicitly infers the unobservable fault context (e.g. joint failure) and rapidly adjusts its locomotion policy, serving as an active recovery mechanism against unexpected degradation. However, meta-learning often relies on the diversity and representativeness of training tasks: if the distribution of training tasks is insufficient, the model will still face challenges when new tasks come. Furthermore, the meta-training process is computationally expensive, requiring repeated optimization at the 'task level'. Additionally, meta-learning models may experience negative transfer, where interference between tasks prevents the optimal initial point for all tasks [54].

Representation learning has been a central focus and is also one of the key factors for achieving successful domain generalization. Typically, the prediction function h is decomposed as $h = f \circ g$, where g is the representation learning function and f is the classifier function. The core objective is to learn a robust representation function g, with its optimization goal as:

$$\min_{f,g} \mathbb{E}_{x,y} \ell\left(\mathrm{f}\left(\mathrm{g}\left(x\right)\right), \mathrm{y}\right) + \lambda \ell_{\mathrm{reg}} \tag{7}$$

where ℓ_{reg} represents the regularization term, and λ is a balancing parameter.

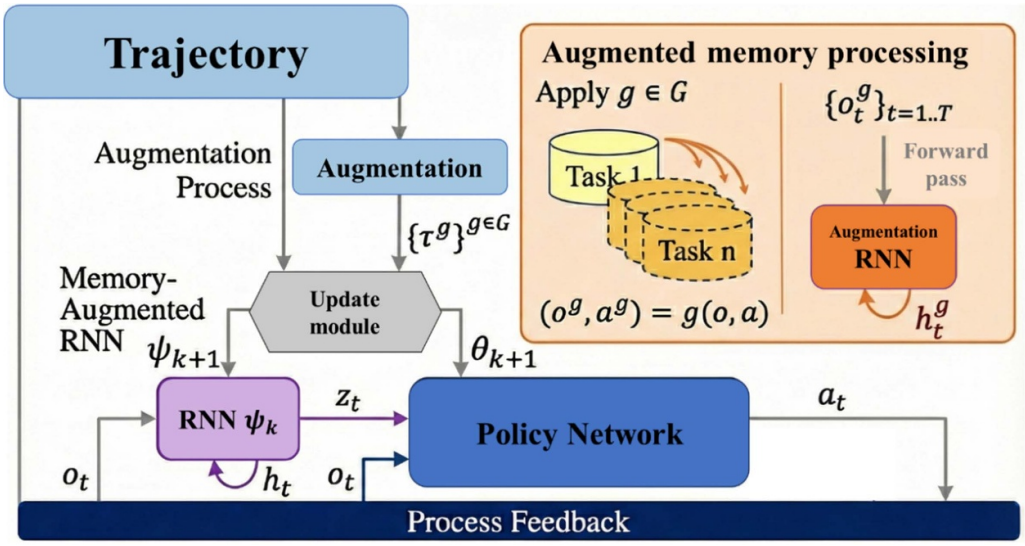


Figure 5. Structure of the memory augmentation strategy.

David *et al* [55] took a theoretical approach and demonstrated that if the feature representation remains invariant, the representation exhibits strong generalization for effective transferring to unseen domains. Blanchard *et al* [56] first introduced kernel methods into domain generalization, using semi-definite kernels to learn invariant kernels. Methods such as KDDRL [57] and CDIG [58] aim to minimize the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$ by aligning feature distributions. KDDRL employs a multi-teacher distillation framework to filter out domain-specific 'style' noise, effectively enforcing invariance. Similarly, CDIG utilizes contrastive learning to pull features of homologous signals closer. However, a fundamental theoretical limitation of strictly invariant representations is the potential loss of semantic discriminability. As discussed in recent studies [59], aggressively aligning distributions may force the model to discard domain-specific features that are actually causally linked to the label (spurious invariance). Furthermore, as noted in the survey by Khoee *et al* [60], these methods struggle when the 'invariant' assumption is violated, i.e. when data distributions differ substantially.

The choice between representation learning and meta-learning depends on the nature of the target shift. Representation learning is preferable when the task remains constant but the environment changes (e.g. autonomous driving in different cities), as it focuses on invariant feature extraction. Conversely, meta-learning is the superior choice for multi-task scenarios where the agent must adapt to entirely new tasks or dynamics with minimal data. However, if computational resources are constrained during the inference phase, representation learning is more suitable, as meta-learning typically requires computationally expensive gradient updates for adaptation.

C. Advanced learning paradigms

Training data diversity and task richness can benefit models' generalization potential. Based on this idea, some advanced learning paradigms such as multi-task learning and large-scale pre-training have become widely adopted approaches.

Multi-task learning improves generalization by jointly training on multiple tasks with shared parameters. This paradigm supports task reconfigurability by fostering versatile representations that allow the system to switch functions reliably without catastrophic performance drops. For example, the authors in [61] treat object recognition in different visual styles as multiple separate tasks. By sharing a core feature-extraction network, the model learns general features to better recognize objects in new domains. In addition, **large-scale pre-trained models** have provided a new perspective on improving IASs generalization. These models leverage vast, diverse datasets for self-supervised pre-training, enabling strong generalization and adaptation to new tasks via fine-tuning [42]. The cross-domain big data compensate for the lack of training data in single tasks, endowing the model with broader knowledge and enhanced generalization. In the context of system reliability, these foundation models offer a form of informational redundancy. When specific local sensor data is sparse or ambiguous (a potential failure point), the model's vast, pre-trained generalized prior acts as a fallback or redundant information source, ensuring decision-making continuity in edge cases absent from the task-specific training data. Google's Robotics Transformer 1 (RT-1) [62] uses over 700 robotic manipulation tasks, 130,000 real-world operation trajectories, enabling significant zero-shot generalization. DeepMind's RT-2 [63] integrates web and robotic data for enhanced semantic reasoning. Experiments showed that RT-2 exhibited remarkable zero-shot generalization on entirely new objects and instructions. The results indicate that its potential to overcome traditional models' generalization limits in complex, open-world scenarios. Similar explorations are also underway in autonomous driving. Wu *et al* [42] reviewed the potential of foundation models, noting that large models significantly enhance scene understanding and

reasoning generalization by leveraging extensive pre-training. However, these models demand significant computational resources, encompassing data collection, storage, and computation. Moreover, these large models severely lack interpretability, with decision-making processes difficult to interpret, which could introduce new risks in safety-critical IASs applications.

These advanced paradigms represent the current state-of-the-art in zero-shot generalization, leveraging massive pre-training to cover diverse edge cases. The critical trade-off here lies in 'safety assurance'. While performance is high, these 'black-box' models severely lack interpretability, making it difficult to trace the root cause of failures. Consequently, while they are applicable to high-level reasoning and perception tasks, their deployment in safety-critical decision-making requires additional safety guardrails compared to traditional methods.

Overall, synthesizing these strategies from the perspective of reliability assurance, distinctions emerge regarding their verifiability and impact on system safety boundaries. Data-centric approaches, particularly domain randomization, offer a deterministic advantage for reliability assessment by enabling engineers to explicitly define and stress-test the ODD. This capability allows for the verification of system performance against prescribed perturbation limits, to align with the established safety standards. In stark contrast, while foundation models provide extensive knowledge redundancy for open-world scenarios, their 'black-box' nature introduces significant epistemic uncertainty (EU). The opacity of their decision-making processes impedes the traceability of specific failure modes, posing severe challenges for certification in safety-critical applications. Meanwhile, model-centric strategies such as meta-learning introduce a dynamic layer of reliability through adaptive fault tolerance; however, their contribution to system resilience is fundamentally contingent upon guaranteeing the algorithm stability of the adaptation process, lest the mechanism itself induces secondary failures during runtime operation.

2.4. Robustness

Robustness is a cornerstone of reliable intelligent systems, denoting their ability to sustain stable performance and functionality in the presence of uncertainty, noise, and environmental perturbations [64]. As autonomous systems increasingly operate in open-ended, unstructured environments, their susceptibility to adversarial perturbations, noisy inputs, and domain shifts poses fundamental challenges to achieving trustworthy operation. Empirical evidence has consistently underscored these vulnerabilities. For example, Ko *et al* [65] revealed that small input perturbations propagate exponentially through time steps, degrading long-term predictions by 25%–40%. Similarly, Goodfellow *et al* [66] demonstrated that adversarial examples exploit the linear nature of intelligent models, causing significant misclassifications even with minimal input changes. These observations highlight the urgent need to address robustness gaps in IASs.

2.4.1. Definition and scope: from perturbation resilience to uncertainty management. Robustness refers to the capacity of IASs to maintain consistent, reliable performance and functionality despite exposure to uncertainties, noise, adversarial inputs, and environmental variations [24]. Formally, it can be conceptualized as a system's resilience to perturbations that deviate from nominal operating conditions, ensuring minimal degradation in key metrics such as accuracy, precision, and mission success rate. Let $x \in X$ be the input, $y \in Y$ be the target output, and $\mathcal{D}$ be the underlying data distribution. We define Δ as the set of allowable perturbations. The robustness objective is mathematically formulated as finding the optimal parameters θ^* that satisfy:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\Delta} L\left(f_\theta\left(x+\delta\right), y\right) \right] \tag{8}$$

where L is the loss function; the inner maximization represents the perturbations or attacks that quantify the system's vulnerability limit; while the outer expectation ensures the system remains resilient across the operational domain.

In the prediction and decision-making processes of IASs, robustness relies heavily on the management of uncertainties, which typically stem from two distinct sources, namely the aleatoric and epistemic uncertainties [67]. The aleatoric uncertainty (AU) primarily arises from the inherent randomness and noise in the environment or sensor data. For instance, the LIDAR point cloud noise during heavy rain is often irreducible even with more data. Conversely, the EU stems from the lack of knowledge within the model itself, often caused by limited training data or distribution shifts. For example, a self-driving car trained only on sunny highways may exhibit high EU when encountering a snowy urban intersection. Quantifying these uncertainties allows the system to distinguish between 'noisy data' and 'unknown scenarios', providing a theoretical basis for fail-safe decisions [68]. In contrast to traditional engineering approaches, which focus primarily on hardware redundancy and logical fault tolerance, the concept of robustness in IASs extends these principles to the algorithmic layer. It serves as a critical reliability assessment dimension, quantifying the system's ability to operate within safe boundaries under uncertainties. In this context, improving robustness is equivalent to minimizing the probability of system failure caused by factors such as unexpected domain shifts or malicious perturbations, thereby directly contributing to the overall mission reliability.

2.4.2. Mechanisms for robustness improvement. In the pursuit of improving the reliability and robustness of IASs, significant advancements have been made, covering various strategies including classical control methods, data augmentation, adversarial training, and RL, to improve the robustness of these systems and their foundational models. Table 3 summarizes different representative methods for robustness improvement.

Table 3. Representative methods for improving robustness in IASs.

Approach	Representative works	Key contributions	Relative Effectiveness & strengths	Limitations & trade-offs
Classical robust control	Li *et al,* Badings, Wei *et al* [69–71]	**Model-based**: Adaptive controller, Interval MDPs, PAC bounds	• Rigorous theoretical guarantees and stability for known dynamics; • Effective for low-level stabilization	• Relies on precise mathematical abstraction; • Difficult to apply and scale
Adversarial training & data augmentation	Shu *et al*, Zhang *et al,* Unal *et al* [72–74]	**Optimization-based:** Min-max optimization, PGD, differentiable augmentation	• Serves as 'virtual fault injection', reducing sensitivity to perturbations and sensor noise	• May degrade performance on clean data; • Computationally expensive for training
Adversarial reinforcement learning	Pinto *et al,* He *et al* [75–77]	**Game-theoretic**: Protagonist vs adversary, actor-critic RL training	• Enables agents to learn resilient policies against disturbances and errors	• Can be hard to converge; • Potential for over-conservative under strong adversary
Attention mechanisms & fusion	Almalioglu *et al*, Zhou *et al*, Dahal *et al* [78–80]	**Architecture-based:** Reliability masks, Self-attention, token clustering	• Acts as intelligent redundancy; • Dynamically reallocates trust from noisy/faulty sensors to reliable ones	• Complexity: e.g. quadratic complexity of self-attention; • Limited robustness under coherent attacks

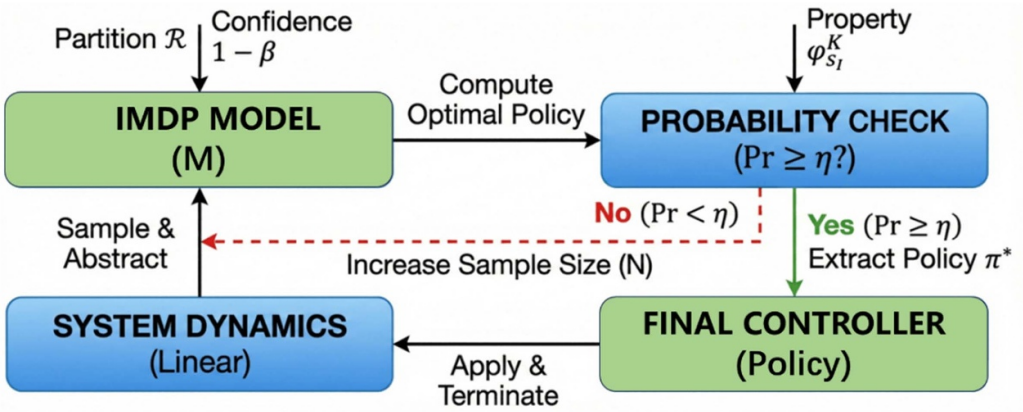


Figure 6. Conceptual illustration of the iterative sampling-based robust control.

1) Robust control methods

The classical **robust control methods** have played a fundamental role in ensuring the reliability of IASs by mitigating the effects of uncertainties and noise. Li *et al* [69] addressed the uncertainty and noise in autonomous vehicles path tracking by introducing an adaptive robust controller, which utilized a linear quadratic regulator for stabilizing the nominal system. An adaptive control law was also designed to suppress uncertainties and measurement noise. Badings *et al* [70] introduced a sampling-based robust control approach as shown in figure 6. The method leveraged finite noise samples to abstract continuous dynamics into interval Markov decision processes (MDPs), employing the scenario approach to compute probably approximately correct (PAC) bounds on transition probabilities. This method aligns with reliability assessment frameworks by abstracting continuous dynamic uncertainties into probabilistic guarantees. It ensures that the UAV motion planning remains within a safe reliability margin even under stochastic noise, effectively acting as a proactive fault-tolerance mechanism. Similarly, Wei *et al* [71] introduced a

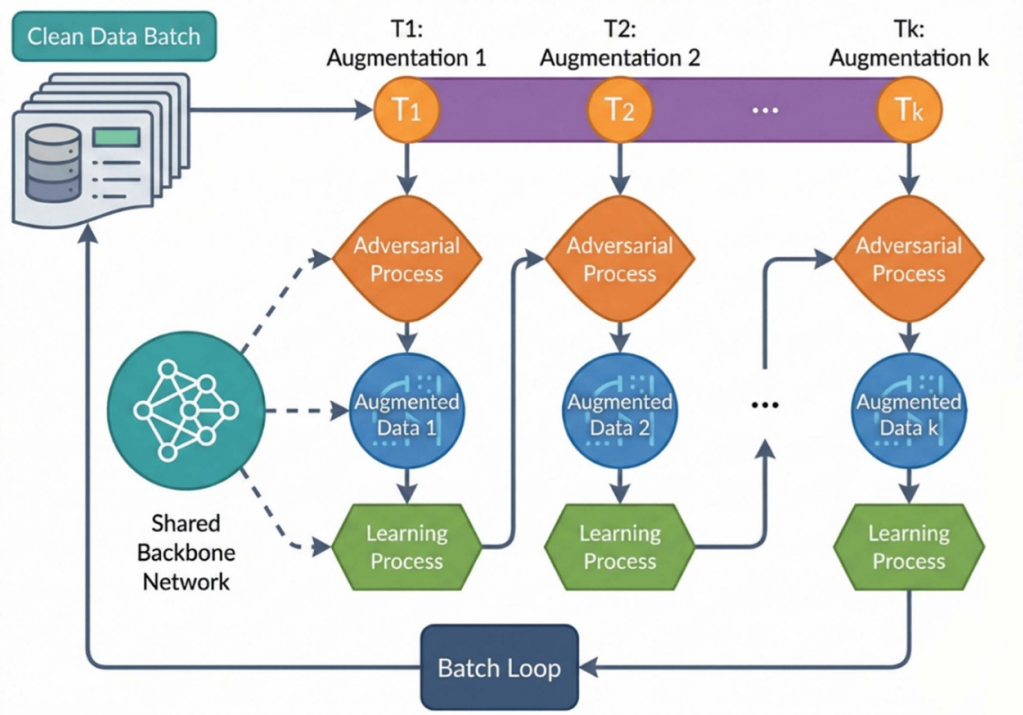


Figure 7. Conceptual illustration of the adversarial data augmentation process.

point-based MDP to tackle the uncertainties in single-lane autonomous driving. Their approach integrated Monte Carlo sampling, probabilistic reasoning with an acceleration prediction model, which enabled to account for sensor noise, and perception of constraints and surroundings, improving robustness in dynamic traffic scenarios.

Classical robust control approaches are preferable for safety-critical control subsystems where the system dynamics can be explicitly modeled. Their primary advantage lies in providing deterministic safety margins and PAC bounds, which are currently difficult to derive for deep learning models. However, its applicability is often limited by the complexity of modeling high-dimensional sensory data. These methods may struggle to generalize in unstructured environments where precise mathematical abstraction of uncertainty is infeasible.

2) Adversarial training techniques

Another primary approach for ensuring the robustness of systems involves **adversarial training techniques**, which aim to make systems more resilient to input perturbations, sensor noise and adversarial attacks [81]. For instance, several studies have focused on adversarial data augmentation, which involves generating adversarial dataset to simulate real-world disturbances and enrich training datasets. The process can be characterized as in figure 7. From a reliability engineering perspective, the adversarial training and data augmentation function as extensive 'virtual fault injection' campaigns. By exposing the model to worst-case perturbations, these methods reduce the system's sensitivity to input variations, effectively acting as a robustness enhancement strategy for the systems. For instance, Shu *et al* [72] proposed an adversarial differentiable data augmentation method to improve the robustness of vision-based control tasks in autonomous vehicles. By formulating image degradations in a differentiable way, they then used the projected gradient descent (PGD) method to find the worst-case augmentation parameters, thereby improving the neural networks' robustness against image corruptions in the learning to steer task. Zhang *et al* [73] extended adversarial approaches to robust trajectory prediction models, showing that perturbing vehicle trajectories can lead to significant prediction errors. They then designed mitigation techniques such as data augmentation and trajectory smoothing methods based on convolution and SVM, to reduce prediction errors under adversarial conditions. Similarly, Unal *et al* [74] propose an adversarial test set generation method to enhance robustness in autonomous systems under data uncertainty and adversarial attacks. The approach adopted a non-dominated sorting genetic algorithm to generate highly uncertain test data. The framework incorporates dropout layers and UQ to improve model resilience. Furthermore, Madry *et al* [82] formulated adversarial robustness of neural networks as a min-max optimization problem, unifying attack generation and defense training. The PGD was employed to solve the inner maximization for generating effective adversarial examples.

By exposing models to worst-case perturbations, these methods significantly improve resilience against sensory noise and adversarial attacks, a domain where classical control is typically inapplicable. Besides, adversarial training functions as a static, offline defense mechanism, embedding robustness into the model weights. Thus, it serves as a dominant strategy for high-dimensional perception tasks (e.g. vision).

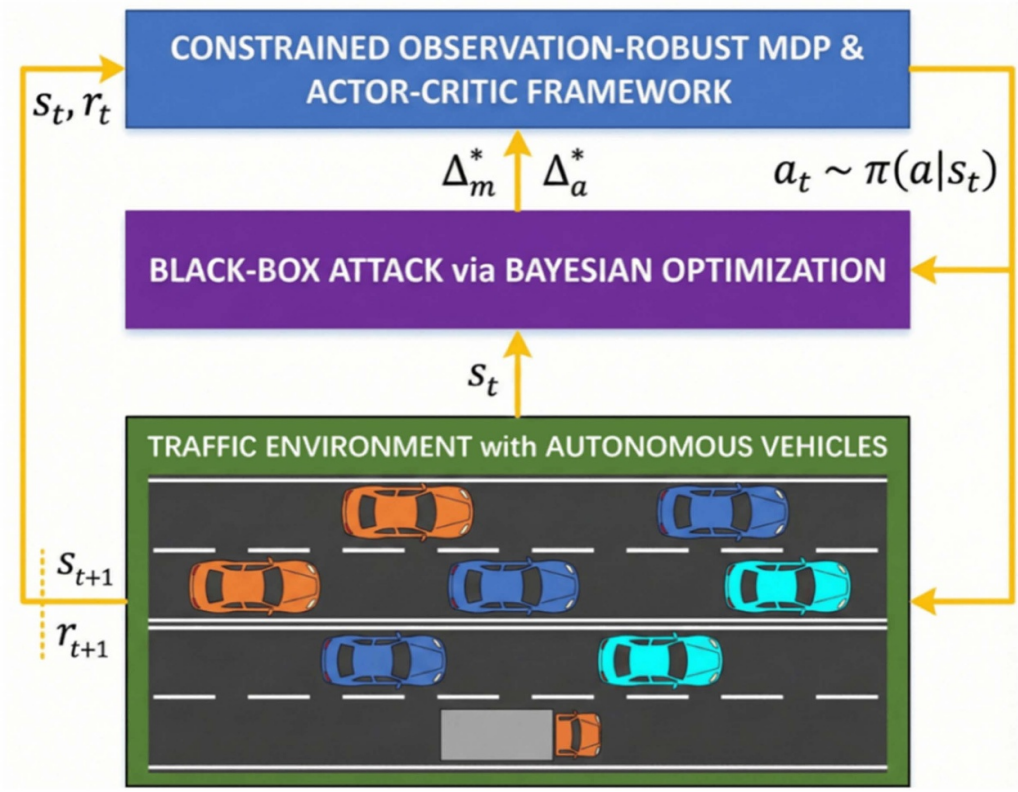


Figure 8. Conceptual illustration of the robust lane change decision-making approach.

Nevertheless, its relative effectiveness comes at a cost: the iterative gradient calculations required for attack generation act as a computationally intensive 'virtual fault injection', substantially adding to the training time. More critically, it introduces a robustness-accuracy trade-off, where the min-max optimization may enforce decision boundaries to accommodate worst-case deviations, thereby compromising its performance on nominal, clean data.

3) Adversarial RL

Moreover, Pinto *et al* [75] introduced **robust adversarial RL** for enhancing policy robustness in RL agents facing modeling errors and environmental uncertainties. The framework jointly trained a protagonist agent and an adversarial agent that applied destabilizing forces, framing the problem as a zero-sum Markov game. The adversary learns optimal disturbance policies to thwart the protagonist, forcing it to develop resilience against perturbations. Similarly, He *et al* have extensively explored the adversarial RL approach to improve autonomous systems robustness. For example, in [76], they model environmental disturbances as an adversarial agent and employ white-box attacks to simulate perturbations. A constrained adversarial actor-critic algorithm was then proposed for on-ramp merging, ensuring variations in the policy and action-value function remain within acceptable bounds. Additionally, an observation adversarial RL framework for robust lane-change decisions was developed in [77]. It employs a constrained MDP to formalize decision-making under perturbations and uses Bayesian optimization for black-box attacks to generate adversarial observations. An observation-robust actor-critic algorithm, as presented in figure 8, minimizes policy variations while optimizing lane-change behaviors. Furthermore, the authors [83] address policy robustness by training an agent to generate optimal perturbations for both observed states and environmental dynamics, enabling the system to learn to resist uncertainties. A safety mask based on the responsibility-sensitive safety model further ensured collision-free policies during training and testing.

Conceptually, the approach bridges the gap between control theory and learning by framing robustness as a zero-sum Markov game. While this formulation allows policies to proactively adapt to worst-case dynamic uncertainties, it faces a critical robustness-performance trade-off. The unconstrained maximization of adversarial or worst-case loss may drive agents toward excessive conservatism, where nominal performance is sacrificed to survive extreme perturbations.

4) Attention mechanisms and fusion

The integration of **attention mechanisms** has also proven its effectiveness in improving robustness, which functions as an intelligent redundancy management strategy. While traditional redundancy relies on voting logic among identical sensors, attention-based fusion allows systems to dynamically assess the trustworthiness of heterogeneous sensors. For instance, the 'reliability masks' proposed by Almalioglu *et al* [78] allow the IASs to dynamically block unreliable sensor streams (e.g. faulty or noisy inputs) and reallocate trust to functional modalities. This mimics the fail-operational logic,

ensuring that partial sensor degradation does not lead to total system failure. Beyond sensor fusion, self-attention enhances the intrinsic robustness of perception models. Zhou *et al* [79] investigated the role of self-attention in the robustness of vision transformers (ViTs) against input corruptions. The authors identify that self-attention promotes mid-level representations through visual token clustering, which correlates with improved robustness. Dahal *et al* [80] presented RobustStateNet, which used RNN to predict vehicle motion, alongside a Kalman filter-like structure to update the predicted state with global positioning data. The system uses a masking mechanism to adaptively weigh and selectively fuse RNN outputs based on feature reliability, enabling it to adapt to measurement uncertainties and noise, particularly in failure-prone scenarios. Zhou *et al* further [79] interpret the above phenomenon via the information bottleneck (IB) principle, suggesting that self-attention optimizes IB objectives by compressing irrelevant information while preserving task-relevant features and, thus, improve the robustness.

Unlike adversarial training, which embeds robustness into model weights via offline optimization, attention mechanisms and fusion strategies operate dynamically at runtime. This offers a distinct advantage in terms of computational efficiency and adaptability. By actively filtering out unreliable sensor streams via reliability masks or attention weights, these methods function as an interpretable, resource-efficient redundancy management strategy. However, this approach introduces specific limitations. The calculation of attention maps, particularly the quadratic complexity of self-attention, imposes a significant computational burden, potentially hindering deployment on resource-constrained devices. Furthermore, while attention acts as a dynamic gatekeeper, it does not inherently guarantee immunity against coherent, semantically disguised attacks.

2.4.3. UQ for robust decision-making. While the aforementioned techniques enhance resilience against specific threats, UQ has emerged as a pivotal strategy for enhancing the robustness of IASs by providing a structured means to measure, propagate, and mitigate uncertainties in dynamic environments [84]. From a reliability engineering perspective, UQ transitions IASs from reactive fault tolerance to proactive risk management, enabling to quantify the confidence in their perceptions, predictions, and decisions, and to adapt accordingly. Unlike traditional robustness methods that focus on worst-case perturbations or redundancy, UQ integrates probabilistic reasoning to bound epistemic and aleatoric uncertainties. This 'confidence calibration' layer significantly reduces the probability of unsafe actions in safety-critical domains such as autonomous driving, robotics, and UASs [85].

1) Types of uncertainty in IASs

As mentioned before, uncertainties in IASs are broadly categorized into AU and EU. Recent surveys highlight the interplay between these types in IASs, where AU predominates in sensor distortions, and EU arises in novel situations [40]. Shao *et al* [86] further refine AU into short-term (e.g. immediate sensor perturbations) and long-term (e.g. multimodal trajectory predictions) components, demonstrating their propagation from prediction to planning stages. Critically, accurate separation of AU and EU enables targeted mitigation. AU requires stochastic modeling to avoid overconfidence, while EU demands detection mechanisms to trigger fail-safe modes, akin to probabilistic safety assessments in reliability engineering [87].

2) Methods for UQ in IASs

State-of-the-art UQ methods integrate principles from probabilistic ML and control theory, providing essential tools for estimating and propagating uncertainties across system components. The current UQ methods can be broadly divided into three categories: Bayesian methods, ensemble methods, and single-network deterministic methods.

Bayesian methods treat model parameters as distributions, enabling posterior inference to capture model uncertainty. BNNs are a prominent example, with variational inference or Monte Carlo Dropout (MC Dropout) used for scalable approximation [88]. For example, Gal and Ghahramani [89] initially established a theoretical link between dropout training and approximate Bayesian inference in deep Gaussian processes, where dropout is retained during inference to perform multiple stochastic forward passes. This technique allows for the estimation of the predictive posterior distribution, thus yielding both predictive mean and uncertainty. For autonomous systems, Kendall and Gal [90] extended Bayesian deep learning to perception tasks. Their framework demonstrated that while AU captures systematic noise, modeling EU is critical for safety-critical applications to detect OOD examples that the model has never seen before. More recently, Franchi *et al* [91] propose the Adaptable BNN, which employs BNN adaptation layers to effectively estimates the posterior distribution around the local minimum of a pre-trained model, thereby converting deterministic DNNs into BNNs for UQ.

Ensemble methods, on the other hand, aggregate predictions from multiple models to quantify uncertainty via output variance [92]. Deep ensembles have become a gold standard owing to their simplicity and strong empirical performance. Tang *et al* [93] combine LSTM with deep ensembles to estimate both epistemic and AU of surrounding vehicles' future trajectories. The uncertainty-aware potential field is fed into model predictive control for uncertainty-aware decision-making, improving safety in lane-changes. Shao *et al* [86] introduce a comprehensive uncertainty management framework, employing deep ensembles to quantify EU, alongside Gaussian mixture models for AU. Their method integrates the uncertainties via tailored risk models and a two-stage training strategy. Extensive evaluation on perception-constrained scenarios shows superior handling of complex traffic interactions compared with deterministic baselines.

While the Bayesian and Ensemble methods provide high-quality estimates, their practicality is often limited by significant memory and inference costs. Consequently, **single deterministic methods** have emerged as a promising solution, enabling reliable UQ within a single forward pass. Van Amersfoort *et al* [94] proposed deterministic UQ (DUQ). By calculating the distance between a feature vector and learned class centroids, DUQ measures uncertainty as the distance to the closest centroid. Crucially, it introduces a penalty to regularize the Jacobian, ensuring matched performance to deep ensembles. Building on the concept of distance awareness, Liu *et al* [95] proposed spectral-normalized neural Gaussian process, which combine spectral normalization of weight matrices with a Gaussian process output layer. This architecture allows to estimate predictive uncertainty via a single forward pass, achieving competitive performance while maintaining simplicity. More recently, Mukhoti *et al* [96] introduced Deep Deterministic Uncertainty (DDU). The method fits a Gaussian discriminant analysis model to the feature space post-training to estimate EU, while using the Softmax entropy for AU. This approach effectively disentangles the two types of uncertainty and has shown state-of-the-art performance on OOD benchmarks relevant to safety-critical applications.

To sum up, these robustness paradigms offer distinct contributions to the reliability assurance landscape. Classical control strategies provide the deterministic underpinnings necessary for certification by offering provable safety margins (e.g. PAC bounds), rendering them indispensable for enforcing absolute stability constraints. However, their limitations in processing high-dimensional sensory data necessitate the integration of adversarial training, which serves as an empirical defense to strengthen IASs against worst-case statistical deviations, despite lacking theoretical guarantees. Complementing these static defenses, attention-based fusion enhances operational fault tolerance by dynamically reallocating trust among redundant modalities during runtime degradation. Ultimately, UQ functions as the critical translation layer for risk management. By disentangling aleatoric noise from epistemic ignorance, it transforms predictive uncertainty into actionable reliability metrics, thereby enabling the implementation of fail-safe triggers essential for safety-critical deployment.

2.5. Explainability

Contemporary IASs often rely on advanced AI models with high-dimensional inputs, non-linear reasoning, and opaque representations, complicating the traceability of output generation [97]. The rapid adoption of IASs has heightened the need for interpretable and trustworthy decision-making processes. Explainability addresses this by enabling IASs to articulate the rationale behind their perceptions, predictions, and actions in a human-accessible manner. It enhances transparency, supports regulatory compliance, facilitates system debugging, and fosters effective human–machine collaboration. As regulatory and societal expectations evolve, robust explainability mechanisms are essential for ensuring functional safety and public trust [98].

2.5.1. Definition and significance of IASs explainability.

Contemporary IASs often rely on advanced AI models characterized by high-dimensional inputs, non-linear reasoning, and opaque representations, which complicates the traceability of output generation. Given these complexities, explainability in IASs is defined as a system's ability to present the reasoning behind its decisions, actions, or predictions in a human-understandable way [99–101]. This capability is crucial for transforming the 'black box' nature of deep learning algorithms into transparent processes where users, developers, and regulators can trace information processing and comprehend specific responses. For instance, in autonomous driving, explainability clarifies which sensor inputs, environmental cues, or decision rules drive actions like lane changes [102, 103]. Likewise, in service robotics, explainability may involve revealing how object recognition processes and task prioritization strategies shape the action sequences [104, 105].

The significance of explainability extends beyond mere transparency; it is a fundamental pillar for ensuring functional safety and public trust as regulatory and societal expectations evolve. From the perspective of reliability engineering, explainability serves as a critical diagnostic interface that mitigates the inherent risks of 'black box' models, where intricate transformations often render decision-making opaque. By elucidating the specific mapping between high-dimensional sensory inputs and control outputs, explainability mechanisms enable the identification of failure modes, allowing engineers to pinpoint the source of erroneous judgments. This capability is essential for validating system integrity, as it allows to verify that decisions are grounded in robust, causal features rather than spurious correlations or data biases, thereby converting uncertainty into actionable insights for fault tolerance and system debugging.

2.5.2. Interpretability frameworks: post-hoc analysis vs ante-hoc design.

To enhance transparency and trust, researchers have developed various explainability techniques, categorized into post-hoc and ante-hoc methods [100]. **Post-hoc** methods seek to interpret and analyze system outputs after training or deployment, providing retrospective insights. In contrast, **ante-hoc** methods incorporate interpretability into the system's architecture or learning, ensuring that the models are inherently more transparent from the outset. Table 4 provides an overview of explainability strategies in IAS, summarizing post-hoc and ante-hoc categories together with their representative methods and core idea.

A. Post-hoc explainability in IASs

Post-hoc explainability analyzes a trained system to reveal how inputs transform into outputs without altering the model. These methods use attribution, visualization, or approximation to uncover decision patterns, aiding bias detection, ensuring appropriate cue focus, and enhancing user trust in deployed IASs where retraining is impractical. Common

Table 4. Comparison of explainability strategies in IAS.

Category	Sub-type	Core idea
Post-hoc	Gradient-based	Use gradients or saliency maps to attribute importance [106, 107]
	Perturbation-based	Alter input or features and observe output changes [108, 109]
	Attention-based	Interpret attention weights as indicators of feature importance [110, 111]
Ante-hoc	Signal-processing-based constraints	Use filterbanks [112], spectral transformers [113], or decomposition layers [114] with physics-aware priors to keep features interpretable
	Sparsity-based constraints	Use sparse priors or regularization [115] to reduce irrelevant features

approaches include gradient-based, perturbation-based, and attention-based methods.

Gradient-based methods attribute model predictions to input features by analyzing how output gradients vary with respect to the inputs. The intuition is that larger gradients indicate features with stronger influence on the decision. In deep neural networks, this process often involves backpropagation. Early methods attribute model predictions to input features by directly computing the first-order derivative of the output with respect to input. For example, Saliency Maps, visualize absolute gradient values to highlight influential regions [106], while Gradient $\times$ Input accounts for feature presence and sensitivity [116]. Formally, for a differentiable intelligent model $f\colon \mathrm{R}^d \to \mathrm{R}^K$ and a target output $f_c(x)$, the derivation of a saliency map can be written as:

$$s_i(x) = \left| \frac{\partial f_c(x)}{\partial x_i} \right|, i = 1, \ldots, d \tag{9}$$

where $s_i(x)$ measures the sensitivity of $f_c(x)$ to feature x_i; Gradient $\times$ Input further defines attributions as $a_i(x) = x_i \frac{\partial f_c(x)}{\partial x_i}$, combining feature magnitude and sensitivity.

These methods are efficient but can produce noisy explanations due to gradient saturation. Advanced techniques like Integrated Gradients [117] accumulate gradients to mitigate local fluctuations. SmoothGrad [118] averages noisy maps to suppress high-frequency noise. Grad-CAM [107] enables coarse localization in visual tasks, refined by Grad-CAM++ [119] via improved localization through weighting adjustment. Recent extensions integrate domain-specific transformations to enhance interpretability in IASs. For example, time–frequency saliency methods with Eigen-CAM [120] highlights critical temporal-spectral regions, supporting decision analysis in autonomous driving radar and drone acoustic sensing. Similarly, hybrid approaches that integrate gradients with Fourier transforms reveal frequency-domain patterns shaping model outputs [121], which are valuable for vibration-based fault diagnosis in autonomous industrial systems. Compared with standard spatial heatmaps, these techniques offer richer interpretive cues for multi-modal or non-visual data.

However, the gradient-based saliency maps can be sensitive to noise and gradient saturation. To mitigate these issues, perturbation-based methods approximate feature importance by directly observing output changes under controlled input modifications. **Perturbation-based methods** attribute importance by altering input or internal features and measuring the resulting effect. If modifying a feature significantly impacts the output, it is deemed critical [108, 109]. Early approaches such as occlusion sensitivity and feature ablation mask specific regions, or channels to evaluate their contributions, while local interpretable model-agnostic explanations (LIMEs) perturbs instances and fit sparse linear surrogates for local interpretability [108]. SHapley Additive exPlanations (SHAP) extends this idea by estimating Shapley values from coalitions of perturbed features, offering theoretical guarantees such as local accuracy and consistency [109]. Formally, consider a feature index set $N = \{1, \ldots, d\}$. For a given IAS input x, the Shapley value of feature i with respect to a target output $f_c(x)$ is defined as:

$$_i(x) = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f\left(x_{S \cup \{i\}}\right) - f(x_S) \right] \tag{10}$$

where x_S denotes the input where only features in subset S are kept and the others are marginalized or replaced by a baseline. The term $f\left(x_{S \cup \{i\}}\right) - f(x_S)$ measures the marginal contribution of feature i on top of coalition S.

Advanced perturbation-based methods like Meaningful Perturbations [122], RISE [123], and extremal perturbations [124] improve robustness, with counterfactual perturbations using generative in-filling to enhance realism [125]. In IASs, the perturbation-based explanations naturally support fault-tolerance analysis by occluding camera patches, masking LiDAR channels, or ablating radar returns which simulate partial sensor failures. The resulting change in action probabilities, $\Delta f_c(x)$, quantifies how much redundancy is available in the remaining modalities. Large $|\Delta f_c(x)|$ indicates a lack of redundancy and suggests that the system may not tolerate the corresponding failure mode in real-world operation. Kim and Canny [102] combined attention mechanisms with causal filtering, where candidate attention regions are masked to verify whether they truly affect steering predictions. Yang *et al* [126] proposed the morphological fragmental perturbation pyramid (MFPP), which perturbs inputs to produce semantic-aligned saliency maps. The method improves interpretive accuracy and efficiency over prior approaches. Figure 9 illustrates the process of MFPP. Puri *et al* [127] proposed Specific and Relevant Feature Attribution for RL agents. The method imposes perturbations on action-specific rewards while penalizing changes to alternative actions, thereby producing interpretable results.

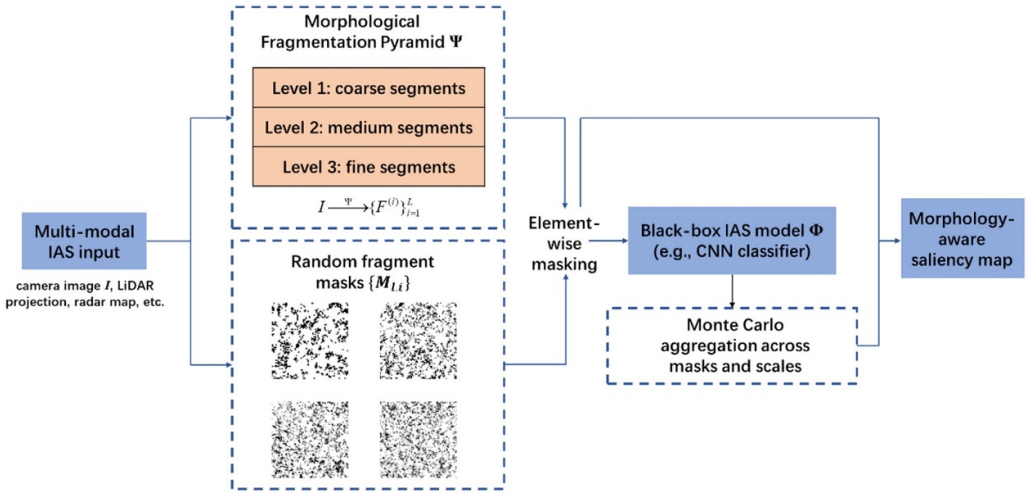


Figure 9. Structure of the MFPP-based perturbation pyramid for IAS explainability and fault analysis.

While perturbation approaches improve faithfulness, their computational cost scales poorly with input dimensionality, especially for high-resolution, multi-sensor IAS data. Architectures with built-in attention mechanisms provide a more scalable alternative. **Attention-based methods** explain model decisions by analyzing the attention weights generated during inference, which indicate the relative importance of components. Basic approaches include raw attention visualization, which projects single-layer or multi-head weights back to the input space [110], and attention rollout, which aggregates weights across layers to capture global dependencies [111]. More advanced variants enhance attribution faithfulness through gradient-based reweighting, extending to non-visual modalities such as LiDAR or time–frequency sensor data. For example, Ichiwara *et al* [128] introduced a modality attention model for robot motion generation, where low-level RNNs process each modality and a high-level RNN fuses them via attention weights. This design reveals which modality drives each task stage, as illustrated in figure 10. Liu *et al* [129] further proposed the Faithfulness Violation Test to evaluate attention-based explanations by examining polarity consistency, i.e. whether highlighted features truly support or suppress predictions. Their results showed that simple methods (e.g. raw attention) are prone to faithfulness violations, whereas gradient-augmented variants reduce errors, with polarity detection ability and model complexity emerging as key determinants of reliability.

Overall, gradient-based methods provide the cheapest explanations, making them attractive for on-board monitoring in real-time IASs. However, their high sensitivity to gradient saturation and adversarial perturbations limits faithfulness, especially under distribution shifts. Perturbation-based techniques, such as LIME or SHAP, generally yield more faithful attributions but incur an order-of-magnitude increase in computational cost, which demonstrate inferiority for high-dimensional multi-sensor inputs and hard real-time constraints. Attention-based explanations are, instead, computationally cheaper and naturally aligned with sequence and multi-modal architectures, yet their reliability in explanations remains debated: raw attention weights can violate basic faithfulness properties without additional reweighting or causal tests.

B. Ante-hoc explainability in IASs

Ante-hoc methods embed interpretability directly into the model during its design and training, by incorporating structural constraints, interpretable components, or domain knowledge, ensuring transparent decision-making throughout operation [130, 131]. These methods produce explanations as a natural byproduct, reducing reliance on external tools and minimizing misleading interpretations. Broadly, ante-hoc methods in IASs follow two major strategies: signal-processing-based constraints [114], and sparsity-based constraints [132].

Signal-processing-based constraints implement explainability by embedding interpretable transforms and physics-aware priors directly into the model design. Foundational techniques include interpretable filterbanks [112], and differentiable spectral transformers [113] that separate trends, harmonics, and transients. Multi-resolution architectures further disentangle temporal and spectral evidence, while spectral or physics-consistency regularizers enforce sparsity, harmonicity, stability, and energy conservation in learned features. Recent advances extend to learnable yet constrained front-ends that retain physical interpretability through structured parameters such as center frequency and bandwidth. For example, the deep morphological convolutional network integrates adaptive morphological filters, enhancing feature discriminability by combining kurtosis-based feature fusion with recalibrated residual learning [114]. Collectively, these approaches demonstrate how integrating domain-specific signal-processing priors can ensure learned representations remain meaningful and transparent.

Early ante-hoc designs mainly relied on hand-crafted, physics-inspired feature extractors. As IAS tasks become

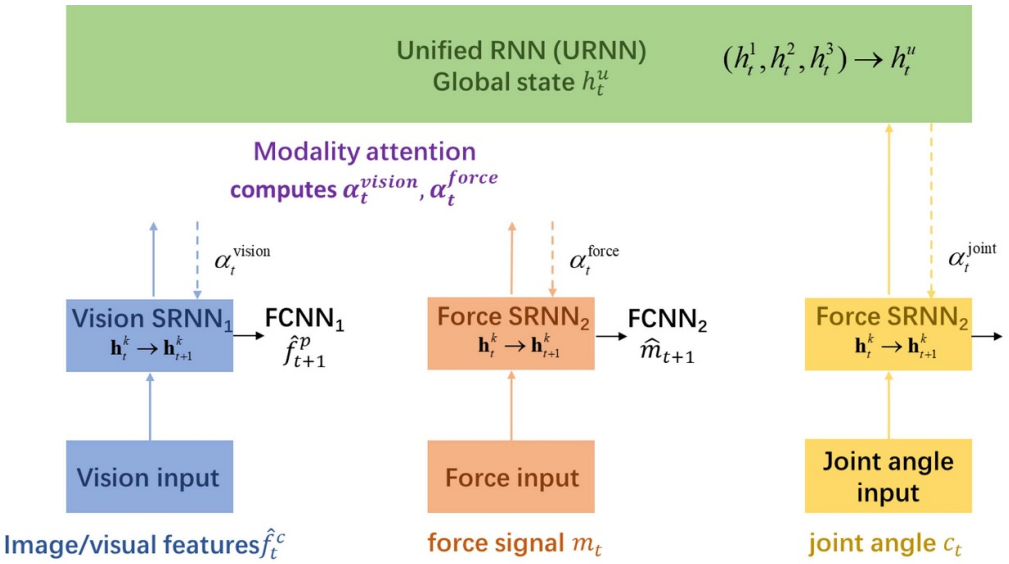


Figure 10. Conceptual illustration of the multi-modal attention as adaptive redundancy allocation in IAS perception.

more complex and data-rich, more flexible architectures with sparsity-based constraints have been proposed to strike a balance between interpretability and predictive power. **Sparsity-based constraints** restrict models to rely on only a small set of salient features or connections, thereby suppressing irrelevant information and highlighting task-relevant patterns. This is typically achieved by sparsity-inducing penalties such as L1 regularization [115], or structured penalties such as Group Lasso [133], which drive weights, channels, or groups toward zero. At the representation level, sparse autoencoders [134] and dictionary learning [135] decompose inputs into clear components, making learned features more interpretable. Building on these principles, recent work has integrated sparsity constraints into intelligent system architectures. For example, Lyons *et al* [132] proposed the deep scale-regularized compound Gaussian network, which models sparse coefficients with a compound Gaussian prior. Each layer corresponds to an iteration of the optimization process, enabling interpretable image reconstruction, thus improving transparency while maintaining accuracy.

Fundamentally, the dichotomy between post-hoc and ante-hoc paradigms determines the feasibility of reliability verification and operational assurance. Post-hoc mechanisms, particularly gradient-based attributions, facilitate retrospective failure root-cause analysis but remain ill-suited for use as runtime safety monitors. Their lack of faithfulness and instability under adversarial noise risk triggering erroneous fallback maneuvers in safety-critical scenarios. While perturbation-based methods offer higher fidelity, their prohibitive computational latency restricts their applicability to offline auditing rather than online fault tolerance, posing a bottleneck for time-critical interventions. In contrast, ante-hoc architectures provide intrinsic verifiability. By aligning decision boundaries with physically interpretable constraints, e.g. signal harmonics or sparsity priors, these models allow for rigorous certification against safety standards prior to deployment. Consequently, the engineering of trustworthy IASs demands a strategic alignment: leveraging ante-hoc designs to enforce deterministic behavior in safety-critical modules, while reserving post-hoc tools for the analysis of complex unconstrained tasks.

2.6. *Synthesis: interdependencies and trade-offs*

Although the above four dimensions are analyzed in isolation, they are intrinsically coupled in guaranteeing the reliability of IASs. From a systemic perspective, these dimensions can be framed as a conceptual hierarchical architecture, bridging the gap between algorithmic properties and system-level reliability. Specifically, ***accuracy*** establishes the baseline for assessment, corresponding to the functional success rate within the nominal ODD. ***Generalization*** and ***robustness*** define the boundaries by quantifying performance degradation under distribution shifts and perturbations, determining the system's safety margins and operational envelopes. Finally, ***explainability*** functions as the audit layer, transforming decisioning process into verifiable logical chains. This serves as a necessary condition for reliability certification, enabling to trace the rationale behind specific actions, facilitate fault diagnosis, and enhance human understanding of the decision-making process. The joint influence of these dimensions dictates the operational safety of the system; a failure in one dimension can trigger a catastrophic collapse of system-level reliability. Specifically, a highly generalizable model that fails on in-distribution tasks, a robust model that sacrifices too much nominal accuracy, or an explainable model whose transparency comes at the cost of unacceptable error rates can offer only limited real-world reliability. Thus, accuracy acts as the common denominator of progress across all dimensions.

While ***generalization*** and ***robustness*** both address performance degradation under distributional shifts, they target distinct aspects and exhibit both overlap and tension. Generalization primarily mitigates EU arising from discrepancies between source and target domains (e.g.

changes in weather) [43]. Robustness, conversely, focuses predominantly on resilience to local perturbations typically within the same nominal data distribution [82]. Empirical studies repeatedly demonstrate that strong generalization does not necessarily imply robustness, nor the reverse. For instance, domain randomization techniques improve sim-to-real transfer and cross-weather generalization in autonomous driving and robotics, yet the trained models may remain vulnerable to adversarial attacks that lie far outside natural distribution shifts [49]. Conversely, standard PGD adversarial training markedly enhances local robustness but can impair generalization to domain shifts (e.g. day-to-night transitions) because the resulting decision boundaries over-emphasize invariance to artificial worst-case perturbations rather than domain gaps.

Another extensively studied trade-off is that between ***robustness*** and ***accuracy*** on clean data. Techniques like adversarial training and uncertainty-aware modeling consistently induce accuracy drops on unperturbed inputs, with reported gaps ranging from 2%–10% in large-scale vision benchmarks [136]. This phenomenon can be attributed to the min–max optimization objective, which forces model to rely on robust but less discriminative features to guarantee performance under worst-case perturbations, thereby sacrificing the ability to exploit subtle patterns present in clean data. In IAS applications, this trade-off manifests concretely as reduced mean average precision in object detection under nominal conditions after adversarial robustness training, or increased trajectory prediction error in clean traffic scenarios following observation-robust RL [76].

Explainability introduces additional multi-dimensional constraints. Ante-hoc interpretability constraints, such as sparsity-inducing regularizers or physics-informed architectures, deliberately restrict the feature space to ensure human-comprehensible decision processes. However, they frequently degrade both accuracy and generalization performance by preventing the model from fitting complex, high-frequency patterns that lack direct physical or causal interpretability [137]. Although post-hoc explanation techniques preserve a model's core accuracy and generalization capabilities, they introduce significant reliability vulnerabilities of their own. Saliency maps and feature attributions can themselves be manipulated by adversarial examples, eroding user trust in high-stakes applications where robustness is essential [138].

A subtler but increasingly recognized interaction exists between ***explainability*** and ***robustness***. Certain inherently interpretable architectures including decision trees, sparse linear models or attention-based models, often exhibit greater robustness to adversarial perturbations because their straightforward form offers fewer degrees of freedom to gradient-based attacks. Conversely, complex deep networks optimized for maximum accuracy or generalization tend to be disproportionately fragile. Recent analyses grounded in the IB principle suggest that enforcing explainability through compression of irrelevant or spurious features can simultaneously improve both robustness and generalization with minimal accuracy penalties [58, 79]. Besides, by revealing whether a decision is based on causal features or spurious correlations, explainability mechanisms allow engineers to detect vulnerable factors that bring improved robustness. For instance, faithfulness violation tests can reveal misaligned attention patterns, guiding refinement of models toward jointly improved transparency and adversarial resilience [129].

In summary, the system-level reliability of IASs is a holistic property that transcends the mere summation of its algorithmic attributes. Rather than pursuing the maximum of individual dimension in isolation, sustainable reliability depends on their strategic synchronization to maintain a stable 'safety margin' across the operational envelope. When effectively harmonized, they create a robust defense-in-depth: accuracy provides the baseline function, generalization and robustness extend that function into uncertain domains, and explainability ensures the entire process remains under human oversight. Consequently, future IAS reliability engineering necessitates an integrative design philosophy that treats these dimensions as a unified system, ensuring operational dependability and functional continuity in complex and dynamic environments.

3. Challenges and future perspectives

3.1. Accuracy

Despite significant advancements in enhancing the accuracy of perception, prediction, and control, several persistent challenges hinder their effective deployment in real-world applications. One primary issue is the accuracy reduction of perception modules under dynamic and constrained conditions, fundamentally linked to limitations in handling non-stationary data distributions [139]. Additionally, although multi-sensor fusion offers potential improvements, its effectiveness is constrained by the challenge of achieving low-latency, deterministic synchronization across sensors with mismatched timestamps and varying sampling rates, all within bounded computational resources [140]. In prediction, a core difficulty lies in accurately modeling the joint evolution of multiple interactive agents. The inherent uncertainty in individual motions and the complex, often discrete nature of interactions is not fully captured by current models, leading to unreliable forecasts in crowded scenarios [141]. Long-term predictions further exacerbate issues, as generative approaches like diffusion models tend to produce physically implausible or overly conservative trajectories beyond short time horizons, indicating a failure to preserve kinematic constraints [142]. Furthermore, in control tasks, RL strategies trained in simulations struggle with the problem of executing guaranteed stable and timely responses on real hardware with limited computing power and unpredictable latency [33].

In future research, first, a primary imperative is the design of a provably consistent state estimation framework for heterogeneous asynchronous sensor suites. This involves developing adaptive fusion algorithms that model and compensate for non-uniform, time-varying sensor latencies, moving beyond post-hoc synchronization, to guarantee estimation accuracy under strict real-time execution budgets [143]. Second, a central challenge in prediction is the integration

of discrete high-level interactive intent with continuous low-level kinematic feasibility within a unified generative model. Solving this requires new architectures that can separately parameterize an agent's strategic decisions and its dynamical constraints, ensuring that long-horizon trajectory distributions remain physically admissible, even in unseen scenarios [144]. In control domains, focusing on robust RL with sim-to-real transfer strategies, such as domain randomization and online adaptation, can address deployment gaps and enhance safety by incorporating negative examples for hazard avoidance [36]. Progress on these well-defined fronts is essential for transitioning IASs from laboratory-validated accuracy to field-deployable reliability.

3.2. Generalization

For IASs, a fundamental challenge to generalization lies in the lack of solid theoretical guarantees for the deep learning models. Current error bounds and learning guarantees are often derived under restrictive or idealized assumptions, which fail to capture the full spectrum of uncertainties encountered in real-world applications [145]. This theoretical gap limits our ability to systematically design algorithms that can adapt to unseen domains or evolving environments. Another persistent obstacle stems from the widespread adoption of the 'closed-world' assumption in existing IAS algorithms, where training and test distributions are presumed to be broadly similar, especially for the data-centric methods. In contrast, real-world environments are inherently open and continually evolving. This introduces a critical time-dependent challenge: novel classes and optimization target may continuously and inevitably emerge [146]. To maintain performance, IASs must navigate the stability-plasticity dilemma, i.e. adapting to new information without overwriting established knowledge. This conflict often leads to catastrophic forgetting, where historical reliability-critical capabilities are eroded during sequential learning. Consequently, systems verified as reliable at deployment (t_0) may suffer from reliability decay, rendering static safety certifications obsolete. Furthermore, the reliance on superficial statistical correlations, rather than causal structures, makes models brittle to environmental changes, constraining their ability to generalize beyond the training distribution [147].

Looking ahead, future research should prioritize the development of theoretically grounded generalization frameworks tailored for 'open-world' environments [148]. This may involve integrating causal inference into representation learning to extract domain-invariant features and promote transferability across heterogeneous tasks [149]. Advancing continual and lifelong learning techniques can help mitigate catastrophic forgetting and support sustained performance as systems encounter novel conditions over time [150]. Additionally, fostering robust multi-modal fusion strategies that explicitly account for cross-modal consistency can enhance generalization by leveraging complementary information across sensors. Recent progress in multimodal perception for autonomous driving, where semantic alignment across visual, LiDAR, and map data improves generalization under rare weather or lighting conditions, exemplifies this potential [151]. These emerging directions offer promising avenues to bridge the gap between empirical success and principled generalization, enabling IASs to operate reliably across diverse and evolving real-world environments.

3.3. Robustness

Although substantial progress has been made in enhancing the robustness of IASs, persistent challenges continue to impede their adoption in critical domains, with the most prominent being the intrinsic trade-off between robustness and accuracy. For example, robust optimization techniques often improve resilience at the expense of reduced performance. The ViTs analyzed in [79] exhibit a 2%–5% drop in accuracy compared to non-robust counterparts. This trade-off can be partially grounded in the min-max optimization framework of adversarial training, which balances worst-case performance against nominal accuracy but frequently results in suboptimal equilibria [136]. A critical gap remains in bridging indicators of robustness performance with system-level reliability metrics. While current works focus on accuracy under perturbation, they often lack a direct mapping to standard reliability metrics such as MTBFs or hazard rates. The challenge lies in translating the statistical loss of a neural network into a deterministic integrity level required for certification in critical domains. Another critical challenge is the high computational load and efficiency associated with robustness-enhancing methods. Techniques such as adversarial training require iterative generation of perturbed examples, often multiplying training time compared to standard methods. This overhead is exacerbated in resource-constrained IASs, where real-time processing is essential, leading to increased energy consumption and latency that can compromise operations [152]. For example, in LSTM-based sequential decision-making, certification processes can extend to hours, making large-scale deployments infeasible [65, 153].

In the future, research should focus on the development of lightweight robust models to mitigate computational overhead while preserving high performance. This could involve model compression techniques, such as pruning and quantization, integrated with adversarial defenses to create efficient architectures for edge devices in IASs. For instance, lightweight end-to-end multimodal models have shown promise in autonomous driving, achieving robust perception with reduced parameters and inference times, paving the way for scalable deployments [154]. Additionally, integrating multi-modal robustness strategies, such as deep learning-based fusion of visual, auditory, and tactile data, can enhance system resilience by leveraging complementary modalities to counteract perturbations in a single input stream. Recent frameworks combining contrastive learning with federated approaches have demonstrated improved robustness in autonomous vehicles, suggesting a path toward distributed decision-making strategies [155]. Emerging technologies like generative AI may also offer transformative potential for IASs robustness, enabling synthetic data generation to simulate diverse adversarial scenarios and refine

decision-making processes. This could address scalability by automating robustness certification and adapting models in real-time to evolving threats, but requires resolving algorithmic challenges in diffusion-based generation, such as ensuring generated samples cover tail-end distributions for rare-event robustness [156]. Moreover, theoretically principled approaches, such as adaptive learning strategies that dynamically balance robustness-accuracy trade-offs, hold promise for optimizing IASs in uncertain environments [157]. Looking ahead for UQ, future directions should prioritize overcoming its integration challenges to meet engineering demands. A key scientific problem is enhancing UQ's robustness itself, as current methods can falter under adversarial perturbations on uncertainty estimates. This necessitates mathematical advancements in meta-uncertainty modeling, such as hierarchical Bayesian frameworks to quantify 'uncertainty of uncertainty' with provable bounds [87].

3.4. Explainability

Current explainability methods for IASs face a tri-lemma of robustness, efficiency, and expressiveness. Post-hoc gradient-based saliency maps are often plagued by saturation and adversarial sensitivity, leading to unstable explanations [103, 158]. While perturbation-based techniques offer greater faithfulness, they incur prohibitive computational costs unsuitable for high-dimensional, multi-modal sensor data [109]. Conversely, ante-hoc designs embed interpretability via rigid structural constraints, which frequently compromise expressive power and generalization [99, 159]. Addressing these limitations requires advancing research along three aligned directions.

The first imperative is to **formalize quantitative robustness guarantees** for explanations under distribution shifts. Current gradient-based attributions lack stability; a fundamental reliability requirement is that if a model's decision remains invariant under a safety margin, its explanation must not diverge significantly [160, 161]. Future work should focus on deriving rigorous bounds to constrain the discrepancy between explanations of perturbed inputs, thereby ensuring that interpretability is resilient to the noise and adversarial shifts inherent in open-world environments.

The second direction addresses the **computational scalability** of methods. To reconcile the high cost of perturbation- and Shapley-style estimators with the strict latency and energy budgets of embedded platforms, where standard Monte Carlo estimators converge slowly, rendering them impractical for real-time multi-modal sensing [162], research must pivot toward efficient approximation. A practical route is to frame explanation as a resource-constrained optimization problem, potentially by distilling a lightweight explainer from a high-fidelity reference model. This ensures that complex multi-modal attributions can be generated within the real-time constraints of autonomous control loops.

Finally, to resolve the trade-off between transparency and performance, the field must evolve toward **hybrid and reliability-aware architectures**. Rather than relying on purely rigid ante-hoc constraints, future designs should synergize interpretable, physics-aware front-ends with flexible deep learning back-ends [163]. A critical challenge may lie in enforcing causal consistency during joint training, for example ensuring that learned attributions align with physical laws such as vibration modes [164]. To validate these advancements, IAS-specific benchmarks must also be established [165]. These evaluation protocols should go beyond visual plausibility to include quantified reliability metrics, such as improved time-to-failure prediction.

4. Conclusion

In this review, we have systematically examined the reliability of IASs, emphasizing their novel technological features such as data-driven learning paradigms, integrated end-to-end architectures, and autonomous task reconfigurability, alongside the associated failure characteristics, including data dependency, limited generalization, and poor interpretability. By framing IAS reliability through the interconnected dimensions of accuracy, generalization, robustness, and explainability (visually summarized in figure 1), we have synthesized state-of-the-art methodologies and empirical insights from diverse domains, ranging from autonomous vehicles to UAVs and robotic platforms. This work distinguishes itself from prior reviews by moving beyond isolated technical analyses to provide a cohesive framework centered on AI-driven decision-making processes, offering a comprehensive roadmap for the development of trustworthy IASs.

Our analysis underscores that significant progress has been achieved in enhancing IAS reliability. Regarding **accuracy**, advancements in multi-sensor fusion, attention-based models, and RL have yielded substantial improvements in perception, prediction, and control, as evidenced by metrics such as MOTA and trajectory errors reduced to sub-meter levels. In terms of **generalization**, data-centric strategies like augmentation and domain randomization, coupled with model-centric approaches such as meta-learning, have facilitated adaptation to OOD scenarios. **Robustness** has been similarly fortified through adversarial training, classical control integration, and UQ, effectively mitigating vulnerabilities to environmental perturbations. Meanwhile, **explainability** techniques spanning post-hoc gradient- and perturbation-based methods to ante-hoc sparsity and signal-processing constraints, have significantly advanced system transparency, thereby fostering user trust and regulatory compliance.

However, despite these advancements, substantial challenges persist that impede the widespread deployment of IASs. In terms of accuracy, perception limits under extreme conditions and the difficulty of modeling complex agent interactions remain critical bottlenecks. Generalization is severely constrained by the prevalent 'closed-world' assumption and a lack of theoretical guarantees, leaving systems vulnerable to catastrophic forgetting in evolving environments. Furthermore, robustness enhancement often incurs a penalty on nominal accuracy and computational efficiency, creating a difficult trade-off for resource-constrained platforms. Similarly, explainability methods currently face a tri-lemma,

struggling to balance robustness, efficiency, and expressiveness simultaneously. To overcome these hurdles, future research directions include leveraging causal inference and lifelong learning for open-world adaptation, developing lightweight models and generative AI strategies to ensure scalable robustness, and establishing reliability-aware architectures that integrate physics-based priors to reconcile transparency with performance.

Ultimately, the reliability of modern IASs is governed not merely by hardware integrity, but increasingly by the logical correctness and adaptability of their intelligent decision-making algorithms. As these systems become integral to safety-critical domains, ensuring their reliability is paramount to preventing failures that could threaten human life and operational safety. This review provides a comprehensive review and roadmap grounded in the four essential algorithmic pillars—accuracy, generalization, robustness, and explainability. By holistically addressing the interdependencies and distinct requirements of these dimensions, the field can advance toward developing autonomous systems that are not only high-performing but fundamentally trustworthy, resilient, and capable of safely navigating the complexities of the real world.

Acknowledgment

This work was supported by the National Natural Science Foundation of China Grant No. U2441271.

ORCID iDs

Jie Liu 0000-0003-0895-7598
Yunxia Chen 0000-0001-9752-8650
Jing Lin 0009-0001-8745-6996

Reference

[1] Budiyono A, Riyanto B and Joelianto E (eds) 2009 *Intelligent Unmanned Systems: Theory and Applications* vol 192 (Springer)
[2] Théron P and Kott A 2019 When autonomous intelligent goodware will fight autonomous intelligent malware: a possible future of cyber defense *MILCOM 2019–2019 IEEE Military Communications Conf. (MILCOM)* (IEEE) pp 1–7
[3] Abhirup R and Akash S 2025 Alphabet's Waymo picks up speed as Tesla robotaxi service expands (available at: www.reuters.com/business/autos-transportation/alphabets-waymo-picks-up-speed-tesla-robotaxi-service-expands-2025-07-15/)
[4] Rainford, M 2025 Baidu partners with uber to deploy apollo go robotaxis globally (available at: https://insidechinaauto.com/2025/07/16/baidu-partners-with-uber-to-deploy-apollo-go-robotaxis-globally/)
[5] Chen J, Sun J and Wang G 2022 From unmanned systems to autonomous intelligent systems *Engineering* **12** 16–19
[6] Chen H, Wen Y, Zhu M, Huang Y, Xiao C, Wei T and Hahn A 2021 From automation system to autonomous system: an architecture perspective *J. Mar. Sci. Eng.* **9** 645
[7] Mani G, Bhargava B, Angin P, Villarreal-Vasquez M, Ulybyshev D and Kobes J Machine learning models to enhance the science of cognitive autonomy *2018 IEEE First Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE)*
[8] Zhang T, Li Q, Zhang C-S, Liang H-W, Li P, Wang T-M, Li S, Zhu Y-L and Wu C 2017 Current trends in the development of intelligent unmanned autonomous systems *Front. Inf. Technol. Electron. Eng.* **18** 68–85
[9] LeCun Y and Bengio Y 1995 Convolutional networks for images, speech, and time series *The Handbook of Brain Theory and Neural Networks* ed M A Arbib (MIT Press) pp 255–8
[10] He H, Gray J, Cangelosi A, Meng Q, McGinnity T M and Mehnen J 2020 The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems *2020 3rd Int. Conf. on Intelligent Robotic and Control Engineering (IRCE)* pp 68–74
[11] Jenihhin M, Reorda M S, Balakrishnan A and Alexandrescu D 2019 Challenges of reliability assessment and enhancement in autonomous systems *2019 IEEE Int. Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)* (IEEE) pp 1–6
[12] Müller M, Müller T, Ashtari Talkhestani B, Marks P, Jazdi N and Weyrich M 2021 Industrial autonomous systems: a survey on definitions, characteristics and abilities *at-Automatisierungstechnik* **69** 3–13
[13] Zhang C, Wang J, Yen G G, Zhao C, Sun Q, Tang Y, Qian F and Kurths J 2020 When autonomous systems meet accuracy and transferability through AI: a survey *Patterns* **1** 100050
[14] Zablocki É, Ben-Younes H, Pérez P and Cord M 2022 Explainability of deep vision-based autonomous driving systems: review and challenges *Int. J. Comput. Vis.* **130** 2425–52
[15] Deng Y, Zhang T, Lou G, Zheng X, Jin J and Han Q-L 2021 Deep learning-based autonomous driving systems: a survey of attacks and defenses *IEEE Trans. Ind. Inform.* **17** 7897–912
[16] Kohli P and Chadha A 2019 Enabling pedestrian safety using computer vision techniques: a case study of the 2018 uber inc. self-driving car crash *Future of Information and Communication Conf.* (Springer) pp 261–79
[17] WikiPedia 2025 Boeing A160 Hummingbird (available at: https://en.wikipedia.org/wiki/Boeing_A160_Hummingbird)
[18] WikiPedia 2025 Reliability Engineering (available at: https://en.wikipedia.org/wiki/Reliability_engineering)
[19] Zhang X, Wang T, Ma L and Mahadevan S 2025 Reliability engineering, risk management, and trustworthiness assurance for AI systems *J. Reliab. Sci. Eng.* **1** 022001
[20] Yu G, Tan G, Huang H, Zhang Z, Chen P, Natella R and Lyu M R 2026 A survey on failure analysis and fault injection in AI systems *ACM Trans. Softw. Eng. Methodol.* **35** 1–42
[21] Osborne M, Lantair J, Shafiq Z, Zhao X, Robu V, Flynn D and Perry J 2019 UAS operators safety and reliability survey: emerging technologies towards the certification of autonomous UAS *2019 4th Int.Conf. on System Reliability and Safety (ICSRS)* (IEEE) pp 203–12
[22] Blood J C, Herbert N W and Wayne M R 2023 Reliability assurance for AI systems *2023 Annual Reliability and Maintainability Symp. (RAMS)* (IEEE) pp 1–6
[23] Olamide K, Kuyoro'Shade E M and Oludele A 2020 Autonomous systems and reliability assessment: a systematic review *Am. J. Artif. Intell.* **4** 30–35
[24] Flammini F, Alcaraz C, Bellini E, Marrone S, Lopez J and Bondavalli A 2022 Towards trustworthy autonomous

systems: taxonomies and future perspectives *IEEE Trans. Emerg. Top. Comput.* **12** 601–14
[25] Dabbabi K and Delleji T 2025 Graph neural network-tracker: a graph neural network-based multi-sensor fusion framework for robust unmanned aerial vehicle tracking *Vis. Comput. Ind. Biomed. Art.* **8** 18
[26] Wang Z, Wang Y, Wu Z, Ma H, Li Z, Qiu H and Li J 2025 CMP: cooperative motion prediction with multi-agent communication *IEEE Robot. Autom. Lett.* **10** 3876–83
[27] Vyacheslav K, Ekaterina C, Egor D and Roman G 2025 Achieving precise and reliable locomotion with differentiable simulation-based system identification (arXiv:2508.04696v1)
[28] Walambe R, Marathe A, Kotecha K, Ghinea G and Doulamis A D 2021 Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions *Comput. Intell. Neurosci.* **2021** 5278820
[29] Musabimana J, Xie Q, Zhou H, Zheng P, Liu H, Ma T and Liu J 2025 A lightweight hybrid model for accurate ammonia prediction in pig houses *Smart Agric. Technol.* **12** 101266
[30] Cheng H, Liu M, Chen L, Broszio H, Sester M and Yang M Y 2023 GATraj: a graph-and attention-based multi-agent trajectory prediction model *ISPRS J. Photogramm. Remote Sens.* **205** 163–75
[31] Lu Y, Xu P, Jiang X, Bashir A K, Gadekallu T R, Wang W and Hu X 2025 Lane change prediction for autonomous driving with transferred trajectory interaction *IEEE Trans. Intell. Transp. Syst.* **26** 4543–56
[32] Zou H, Guo Y, Wei F, Guo D, Li Q and Pirov J 2025 A pedestrian group crossing intention prediction model integrating spatiotemporal features *Sci. Rep.* **15** 20675
[33] Ma B, Liu Z, Zhao W, Yuan J, Long H, Wang X and Yuan Z 2023 Target tracking control of UAV through deep reinforcement learning *IEEE Trans. Intell. Transp. Syst.* **24** 5983–6000
[34] Xia Q, Chen P, Xu G, Sun H, Li L and Yu G 2024 Adaptive path-tracking controller embedded with reinforcement learning and preview model for autonomous driving *IEEE Trans. Veh. Technol.* **74** 3736–50
[35] Ran C, Xie Z, Xie Y, Yin Y and Ye H 2024 A car-following model integrating personalized driving style based on the DER-DDPG deep reinforcement learning algorithm *IEEE Access* **12** 136889–906
[36] Nan J, Zhang R, Yin G, Zhuang W, Zhang Y and Deng W 2025 Safe and interpretable human-like planning with transformer-based deep inverse reinforcement learning for autonomous driving *IEEE Trans. Autom. Sci. Eng.* **22** 12134–46
[37] Lebede N and Nadarajah S 2025 Enhancing autonomous systems with Bayesian neural networks: a probabilistic framework for navigation and decision-making *Front. Built Environ.* **11** 1597255
[38] Yang T, Guo H, Liang H and Yan B 2023 Intelligent combustion control of the hot blast stove: a reinforcement learning approach *Processes* **11** 3140
[39] Liu K 2024 Contact force control of robot polishing system based on vision control algorithm *IEEE Access* **12** 137926–137941
[40] Wang K, Ma Q, Jiang Y and Lu J 2025 Uncertainty quantification for safety of the intended functionality of autonomous driving: a comprehensive survey *IEEE Trans. Instrum. Meas.* **74** 2545122
[41] Filos A, Tigkas P, McAllister R, Rhinehart N, Levine S and Gal Y 2020 Can autonomous vehicles identify, recover from, and adapt to distribution shifts? *Int. Conf. on Machine Learning* (PMLR) pp 3145–53
[42] Wu J *et al* 2024 Prospective role of foundation models in advancing autonomous vehicles *Research* **7** 0399
[43] Wang J, Lan C, Liu C, Ouyang Y, Qin T, Lu W, Chen Y, Zeng W and Yu P S 2022 Generalizing to unseen domains: a survey on domain generalization *IEEE Trans. Knowl. Data Eng.* **35** 8052–72
[44] Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *J. Big. Data* **6** 1–48
[45] Zhou C, He J, Yang S, Xiong X and Chi Y 2025 Domain generalization for the open-set cross-domain diagnosis of a class-imbalanced rod-fastening rotor dataset based on QGAN and aligned reciprocal points adversarial learning *Adv. Eng. Inf.* **68** 103646
[46] Alsafadi F, Furlong A and Wu X 2025 Predicting critical heat flux with uncertainty quantification and domain generalization using conditional variational autoencoders and deep neural networks *Ann. Nucl. Energy* **220** 111502
[47] Wang Y, Li Z, Yu H and Li J 2025 Mixup-based maximum distribution difference selection strategy for domain generalization *Expert Syst. Appl.* **281** 127521
[48] Liang C *et al* 2024 Single domain generalization method for remote sensing image segmentation via category consistency on domain randomization *IEEE Trans. Geosci. Remote Sens.* **62** 1–16
[49] Chen S, Liu G, Zhou Z, Zhang K and Wang J 2023 Robust multi-agent reinforcement learning method based on adversarial domain randomization for real-world dual-UAV cooperation *IEEE Trans. Intell. Veh.* **9** 1615–27
[50] Ma S, Song K, Niu M, Tian H, Wang Y and Yan Y 2024 Feature-based domain disentanglement and randomization: a generalized framework for rail surface defect segmentation in unseen scenarios *Adv. Eng. Inf.* **59** 102274
[51] Finn C, Abbeel P and Levine S 2017 Model-agnostic meta-learning for fast adaptation of deep networks *Int.Conf. on Machine Learning PMLR* pp 1126–35
[52] Yu T, Quillen D, He Z, Julian R, Hausman K, Finn C and Levine S 2020 Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning *Conf. on robot learning PMLR* pp 1094–100
[53] Bao K, Li C, As Y, Krause A and Hutter M 2025 Toward task generalization via memory augmentation in meta-reinforcement learning (arXiv:2502.01521)
[54] Lake B M and Baroni M 2023 Human-like systematic generalization through a meta-learning neural network *Nature* **623** 115–21
[55] Ben-David S, Blitzer J, Crammer K and Pereira F 2006 Analysis of representations for domain adaptation *Advances in Neural Information Processing Systems* vol 19
[56] Blanchard G, Deshmukh A A, Dogan U, Lee G and Scott C 2021 Domain generalization by marginal transfer learning *J. Mach. Learn. Res.* **22** 1–55
[57] Niu Z, Yuan J, Ma X, Xu Y, Liu J, Chen Y-W, Tong R and Lin L 2023 Knowledge distillation-based domain-invariant representation learning for domain generalization *IEEE Trans. Multimedia* **26** 245–55
[58] Xiao X, Zhang J and Xu D 2025 Contrastive domain-invariant generalization for remaining useful life prediction under diverse conditions and fault modes *Reliab. Eng. Syst. Saf.* **253** 110534
[59] Hong Z, Wang Z and Shen L, Yao Y, Huang Z, Chen S, Yang C, Gong M and Liu T 2024 Improving non-transferable representation learning by harnessing content and style *The 12th Int.Conf. on Learning Representations*
[60] Khoee A G, Yu Y and Feldt R 2024 Domain generalization through meta-learning: a survey *Artif. Intell. Rev.* **57** 285

[61] Qi L, Yang H, Shi Y and Geng X 2024 Multimatch: multi-task learning for semi-supervised domain generalization *ACM Trans. Multimedia Comput. Commun. Appl.* **20** 1–21

[62] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, Gopalakrishnan K, Hausman K, Herzog A, Hsu J and Ibarz J 2022 Rt-1: robotics transformer for real-world control at scale (arXiv:2212.06817)

[63] Zitkovich B *et al* 2023 Rt-2: vision-language-action models transfer web knowledge to robotic control *Conf. on Robot Learning PMLR* pp 2165–83

[64] Braiek H B and Khomh F 2025 Machine learning robustness: a primer *Trustworthy AI in Medical Imaging* (Academic Press) pp 37–71

[65] Ko C Y, Lyu Z, Weng L, Daniel L, Wong N and Lin D 2019 POPQORN: quantifying robustness of recurrent neural networks *Int. Conf. on Machine Learning* (PMLR) pp 3468–77

[66] Goodfellow I J, Shlens J and Szegedy C 2014 Explaining and harnessing adversarial examples (arXiv:1412.6572)

[67] Zeng Z, Kang R, Wen M and Zio E 2017 A model-based reliability metric considering aleatory and epistemic uncertainty *IEEE Access* **5** 15505–15

[68] Hüllermeier E and Waegeman W 2021 Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods *Mach. Learn.* **110** 457–506

[69] Li H, Huang J, Yang Z, Hu Z, Yang D and Zhong Z 2022 Adaptive robust path tracking control for autonomous vehicles with measurement noise *Int. J. Robust Nonlinear Control* **32** 7319–35

[70] Badings T S, Abate A, Jansen N, Parker D, Poonawala H A and Stoelinga M 2022 Sampling-based robust control of autonomous systems with non-Gaussian noise *Proc. AAAI Conf. Artif. Intell.* **36** 9669–78

[71] Wei J, Dolan J M, Snider J M and Litkouhi B 2011 A point-based MDP for robust single-lane autonomous driving behavior under uncertainties *2011 IEEE Int. Conf. on Robotics and Automation* (IEEE) pp 2586–92

[72] Shu M, Shen Y, Lin M C and Goldstein T 2021 Adversarial differentiable data augmentation for autonomous systems *2021 IEEE Int. Conf. on Robotics and Automation (ICRA)* (IEEE) pp 14069–75

[73] Zhang Q, Hu S, Sun J, Chen Q A and Mao Z M 2022 On adversarial robustness of trajectory prediction for autonomous vehicles *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 15159–68

[74] Unal D, Catak F O, Houkan M T, Mudassir M and Hammoudeh M 2023 Towards robust autonomous driving systems through adversarial test set generation *ISA Trans.* **132** 69–79

[75] Pinto L, Davidson J, Sukthankar R, Gupta A 2017 Robust adversarial reinforcement learning *Int. Conf. on machine learning* (PMLR) pp 2817–26

[76] He X, Lou B, Yang H and Lv C 2022 Robust decision making for autonomous vehicles at highway on-ramps: a constrained adversarial reinforcement learning approach *IEEE Trans. Intell. Transp. Syst.* **24** 4103–13

[77] He X, Yang H, Hu Z and Lv C 2022 Robust lane change decision making for autonomous vehicles: an observation adversarial reinforcement learning approach *IEEE Trans. Intell. Veh.* **8** 184–93

[78] Almalioglu Y, Turan M, Trigoni N and Markham A 2022 Deep learning-based robust positioning for all-weather autonomous driving *Nat. Mach. Intell.* **4** 749–60

[79] Zhou D, Yu Z, Xie E, Xiao C, Anandkumar A, Feng J and Alvarez J M 2022 Understanding the robustness in vision transformers *Int. Conf. on Machine Learning PMLR* pp 27378–94

[80] Dahal P, Mentasti S, Paparusso L, Arrigoni S and Braghin F 2024 RobustStateNet: robust ego vehicle state estimation for autonomous driving *Robot. Auton. Syst.* **172** 104585

[81] Qian Z, Huang K, Wang Q F and Zhang X-Y 2022 A survey of robust adversarial training in pattern recognition: fundamental, theory, and methodologies *Pattern Recogn.* **131** 108889

[82] Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A 2017 Towards deep learning models resistant to adversarial attacks (arXiv:1706.06083)

[83] He X, Huang W and Lv C 2024 Toward trustworthy decision-making for autonomous vehicles: a robust reinforcement learning approach with safety guarantees *Engineering* **33** 77–89

[84] Wang T *et al* 2025 From aleatoric to epistemic: exploring uncertainty quantification techniques in artificial intelligence (arXiv:2501.03282)

[85] Dutta R G, Guo X and Jin Y 2016 Quantifying trust in autonomous system under uncertainties *2016 29th IEEE Int. System-on-Chip Conf. (SOCC)* (IEEE) pp 362–7

[86] Shao W, Xu J, Cao Z, Wang H and Li J 2025 From prediction to planning: comprehensive uncertainty management in autonomous driving *IEEE Trans. Intell. Transp. Syst.* **26** 16466–80

[87] Wang K, Shen C, Li X and Lu J 2025 Uncertainty quantification for safe and reliable autonomous vehicles: a review of methods and applications *IEEE Trans. Intell. Transp. Syst.* **26** 2880–96

[88] Abdar M *et al* 2021 A review of uncertainty quantification in deep learning: techniques, applications and challenges *Inform. Fusion* **76** 243–97

[89] Gal Y and Ghahramani Z 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning *Int. Conf. on Machine Learning* (PMLR) pp 1050–9

[90] Kendall A and Gal Y 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* vol 30

[91] Franchi G, Laurent O, Leguéry M, Bursuc A, Pilzer A and Yao A 2024 Make me a BNN: a simple strategy for estimating Bayesian uncertainty from pre-trained models *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 12194–204

[92] Rahaman R 2021 Uncertainty quantification and deep ensembles *Advances in Neural Information Processing Systems* vol 34 pp 20063–75

[93] Tang X, Yang K, Wang H, Wu J, Qin Y, Yu W and Cao D 2022 Prediction-uncertainty-aware decision-making for autonomous vehicles *IEEE Trans. Intell. Veh.* **7** 849–62

[94] Van Amersfoort J, Smith L, Teh Y W and Gal Y 2020 Uncertainty estimation using a single deep deterministic neural network *Int. Conf. on Machine Learning* (PMLR) pp 9690–700

[95] Liu J, Lin Z, Padhy S, Tran D, Bedrax Weiss T and Lakshminarayanan B 2020 Simple and principled uncertainty estimation with deterministic deep learning via distance awareness *Advances in Neural Information Processing Systems* vol 33 pp 7498–512

[96] Mukhoti J, Kirsch A, Van Amersfoort J, Torr P H and Gal Y 2023 Deep deterministic uncertainty: a new simple baseline *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 24384–94

[97] Angelov P P, Soares E A, Jiang R, Arnold N I and Atkinson P M 2021 Explainable artificial intelligence: an analytical review *WIREs Data Mining Knowl. Discov.* **11** e1424

[98] Minh D, Wang H X, Li Y F and Nguyen T N 2022 Explainable artificial intelligence: a comprehensive review *Artif. Intell. Rev.* **55** 3503–68

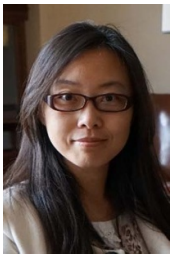
[99] Doshi-Velez F and Kim B 2017 Towards a rigorous science of interpretable machine learning (arXiv:1702.08608)
[100] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F and Pedreschi D 2018 A survey of methods for explaining black box models *ACM Comput. Surv.* **51** 1–42
[101] Samek W, Wiegand T and Müller K R. 2017 Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models (arXiv:1708.08296)
[102] Kim J and Canny J 2017 Interpretable learning for self-driving cars by visualizing causal attention *Proc. IEEE Int. Conf. on computer vision* pp 2942–50
[103] Montavon G, Samek W and Müller K R 2018 Methods for interpreting and understanding deep neural networks *Digit. Signal Process.* **73** 1–15
[104] Anjomshoae S *et al* 2019 Explainable agents and robots: results from a systematic literature review *18th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2019)* (*Montreal, Canada*, *13–17 May 2019*) (International Foundation for Autonomous Agents and Multiagent Systems) pp 1078–88
[105] Miller T 2019 Explanation in artificial intelligence: insights from the social sciences *Artif. Intell.* **267** 1–38
[106] Simonyan K, Vedaldi A and Zisserman A 2013 Deep inside convolutional networks: visualising image classification models and saliency maps (arXiv:1312.6034)
[107] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-cam: visual explanations from deep networks via gradient-based localization *Proc. IEEE Int. Conf. on Computer Vision* pp 618–26
[108] Ribeiro M T, Singh S and Guestrin C 2016 Model-agnostic interpretability of machine learning (arXiv:1606.05386)
[109] Lundberg S M and Lee S I 2017 A unified approach to interpreting model predictions *Advances in Neural Information Processing Systems* vol 30
[110] Chefer H, Gur S and Wolf L 2021 Transformer interpretability beyond attention visualization *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 782–91
[111] Abnar S and Zuidema W 2020 Quantifying attention flow in transformers (arXiv:2005.00928)
[112] Ravanelli M and Bengio Y 2018 Interpretable convolutional filters with sincnet (arXiv:1811.09725)
[113] Tokozume Y and Harada T 2017 Learning environmental sounds with end-to-end convolutional neural network *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 2721–5
[114] Ye Z and Yu J 2021 Deep morphological convolutional network for feature learning of vibration signals and its applications to gearbox fault diagnosis *Mech. Syst. Signal Process.* **161** 107984
[115] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–88
[116] Shrikumar A, Greenside P and Kundaje A 2017 Learning important features through propagating activation differences *Int. Conf. on Machine Learning* (PMLR) pp 3145–53
[117] Sundararajan M, Taly A and Yan Q 2017 Axiomatic attribution for deep networks *Int. Conf. on Machine Learning* (PMLR) pp 3319–28
[118] Smilkov D, Thorat N, Kim B, Viégas F and Wattenberg M 2017 Smoothgrad: removing noise by adding noise (arXiv:1706.03825)
[119] Chattopadhay A, Sarkar A, Howlader P and Balasubramanian V N 2018 Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks *2018 IEEE WINTER Conf. on Applications of Computer Vision (WACV)* (IEEE) pp 839–47
[120] Yu S, Wang M, Pang S, Song L and Qiao S 2022 Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network *Measurement* **196** 111228
[121] Li J, Wang Y, Zi Y and Zhang Z 2021 Whitening-Net: a generalized network to diagnose the faults among different machines and conditions *IEEE Trans. Neural Netw. Learn. Syst.* **33** 5845–58
[122] Fong R C and Vedaldi A 2017 Interpretable explanations of black boxes by meaningful perturbation *Proc. IEEE Int. Conf. on Computer Vision* pp 3429–37
[123] Petsiuk V, Das A and Saenko K. 2018 Rise: randomized input sampling for explanation of black-box models (arXiv:1806.07421)
[124] Fong R, Patrick M and Vedaldi A 2019 Understanding deep networks via extremal perturbations and smooth masks *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 2950–8
[125] Chang C H, Creager E, Goldenberg A and Duvenaud D 2018 Explaining image classifiers by counterfactual generation (arXiv:1807.08024)
[126] Yang Q, Zhu X, Fwu J K, Ye Y, You G and Zhu Y 2021 Mfpp: morphological fragmental perturbation pyramid for black-box model explanations *2020 25th Int. Conf. on Pattern Recognition (ICPR)* (IEEE) pp 1376–83
[127] Puri N *et al* 2019 Explain your move: understanding agent actions using specific and relevant feature attribution (arXiv:1912.12191)
[128] Ichiwara H, Ito H, Yamamoto K, Mori H and Ogata T 2023 Modality attention for prediction-based robot motion generation: improving interpretability and robustness of using multi-modality *IEEE Robot. Autom. Lett.* **8** 8271–8
[129] Liu Y, Li H, Guo Y, Kong C, Li J and Wang S 2022 Rethinking attention-model explainability through faithfulness violation test *Int. Conf. on Machine Learning PMLR* pp 13807–24
[130] Du M, Liu N and Hu X 2019 Techniques for interpretable machine learning *Commun. ACM* **63** 68–77
[131] Puiutta E and Veith E M 2020 Explainable reinforcement learning: a survey *Int. Cross-Domain Conf. for Machine Learning and Knowledge Extraction* (Springer) pp 77–95
[132] Lyons C, Raj R G and Cheney M 2022 A deep compound Gaussian regularized unfolded imaging network *2022 56th Asilomar Conf. on Signals, Systems, and Computers* (IEEE) pp 940–6
[133] Yuan M and Lin Y 2006 Model selection and estimation in regression with grouped variables *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67
[134] Ng A 2011 Sparse autoencoder *CS294A Lecture Notes* vol 72 pp 1–19
[135] Mairal J, Bach F, Ponce J and Sapiro G 2010 Online learning for matrix factorization and sparse coding *J. Mach. Learn. Res.* **11**
[136] Zhang H, Yu Y, Jiao J, Xing E, El Ghaoui L and Jordan M 2019 Theoretically principled trade-off between robustness and accuracy *Int. Conf. on Machine Learning* (PMLR) pp 7472–82
[137] Zheng S, Pan K, Chen Y, Liu J and Zio E 2025 Causality-enhanced system reliability and safety analysis: an overview *Reliab. Eng. Syst. Saf.* **270** 112109
[138] Slack D, Hilgard S, Jia E, Singh S and Lakkaraju H 2020 Fooling lime and shap: adversarial attacks on post hoc explanation methods *Proc. AAAI/ACM Conf. on AI, Ethics, and Society* pp 180–6.
[139] Wang B, Zio E, Chen X, Zhu H, Guo Y and Fan S 2024 Reliability improvement of the dredging perception system: a sensor fault-tolerant strategy *Reliab. Eng. Syst. Saf.* **247** 110134
[140] Liu G, Huang K, Lv X, Sun Y, Li H, Lei X and Shu L 2025 Innovations and refinements in LiDAR odometry and

mapping: a comprehensive review *IEEE/CAA J. Autom. Sinica* **12(6)** 1072–1094
[141] Jiang J, Yan K, Xia X and Yang B 2025 A survey of deep learning-based pedestrian trajectory prediction: challenges and solutions *Sensors* **25** 957
[142] Ni J, Guo Y, Liu Y, Chen R, Lu L and Wu Z 2025 Maskgwm: a generalizable driving world model with video mask reconstruction *Proc. Computer Vision and Pattern Recognition Conf.* pp 22381–91
[143] Liu H, Su T and Guo J 2025 Autonomous driving enhanced: a fusion framework integrating LiDAR point clouds with monovision depth-aware transformers for robust object detection *Eng. Res. Exp.* **7** 015414
[144] Liu D, Wang Y, Liu C, Yuan X, Wang K and Yang C 2024 Scope-free global multi-condition-aware industrial missing data imputation framework via diffusion transformer *IEEE Trans. Knowl. Data Eng.* **36** 6977–88
[145] Sicilia A, Zhao X and Hwang S J 2023 Domain adversarial neural networks for domain generalization: when it works and how to improve *Mach. Learn.* **112** 2685–721
[146] Zhou Z H 2022 Open-environment machine learning *Natl Sci. Rev.* **9** nwac123
[147] Liu J, Zheng S and Wang C 2023 Causal graph attention network with disentangled representations for complex systems fault detection *Reliab. Eng. Syst. Saf.* **235** 109232
[148] Kejriwal M, Kildebeck E, Steininger R and Shrivastava A 2024 Challenges, evaluation and opportunities for open-world learning *Nat. Mach. Intell.* **6** 580–8
[149] Zhang J, Jennings J, Hilmkil A, Pawlowski N, Zhang C and Ma C 2024 Towards causal foundation model: on duality between causal inference and attention *Proc. 41st Int. Conf. on Machine Learning (ICML)*
[150] Wang L, Zhang X, Su H and Zhu J 2024 A comprehensive survey of continual learning: theory, method and application *IEEE Trans. Pattern Anal. Mach. Intell.* **46** 5362–83
[151] Kong L, Xu X, Ren J, Zhang W, Pan L, Chen K, Ooi W T and Liu Z 2025 Multi-modal data-efficient 3D scene understanding for autonomous driving *IEEE Trans. Pattern Anal. Mach. Intell.* **47** 3748–65
[152] Igenewari L S and Okoh O E 2025 Adversarial attacks and defenses in AI systems: challenges, strategies, and future directions *Int. J. Res. Innov. Appl. Sci.* **10** 996–1022
[153] Shivashankar K, Hajj G S A and Martini A 2025 Scalability and maintainability challenges and solutions in machine learning: systematic literature review (arXiv:2504.11079)
[154] Qiao Z, Li H, Cao Z and Liu H X 2025 Lightemma: lightweight end-to-end multimodal model for autonomous driving (arXiv:2505.00284)
[155] Almutairi S and Barnawi A 2025 Enhancing federated learning model adversarial robustness in autonomous vehicles: a lightweight framework with contrastive learning and spatial clustering *Knowl.-Based Syst.* **329** 114253
[156] Andreoni M, Lunardi W T, Lawton G and Thakkar S 2024 Enhancing autonomous system security and resilience with generative AI: a comprehensive survey *IEEE Access* **12** 109470–93
[157] Fawole O and Rawat D 2025 Recent advances in vision transformer robustness against adversarial attacks in traffic sign detection and recognition: a survey *ACM Comput. Surv.* **57** 1–33
[158] Samek W, Montavon G, Vedald A, Hansen L.K. and Müller K.R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* 2019 (Springer)
[159] Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15
[160] Adebayo J, Gilmer J and Muelly M, Goodfellow I, Hardt M and Kim B 2018 Sanity checks for saliency maps *Advances in Neural Information Processing Systems* vol 31
[161] Ghorbani A, Abid A and Zou J 2019 Interpretation of neural networks is fragile *Proc. AAAI Conf. Artif. Intell.* **33** 3681–8
[162] Gamatié A and Wang Y 2024 Explainable AI for embedded systems design: a case study of static redundant NVM memory write prediction (arXiv:2403.04337)
[163] Roscher R, Bohn B, Duarte M F and Garcke J 2020 Explainable machine learning for scientific insights and discoveries *IEEE Access* **8** 42200–16
[164] Raissi M, Perdikaris P and Karniadakis G E 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations *J. Comput. Phys.* **378** 686–707
[165] Sithakoul S, Meftah S and Feutry C 2024 Beexai: benchmark to evaluate explainable ai *World Conf. on Explainable Artificial Intelligence* (Springer) pp 445–68

Jie Liu, Assoc. Prof., received a BS degree in mechanical engineering and an MS degree in physics from Beihang University, Beijing, China, in 2009 and 2012, respectively, and a PhD degree from Centrale Supélec, Gif-sur-Yvette, France, in 2015. He is currently an Associate Professor with the School of Reliability and Systems Engineering, Beihang University, Beijing, China. His research interests include trustworthy fault detection, diagnostics, and prognostics.

Shuwen Zheng received BS and MS degrees from Beihang University, China, in 2021 and 2024, respectively. He is currently pursuing a PhD degree with the School of Reliability and Systems Engineering at Beihang University. His current research interests are artificial intelligence and fault diagnosis.

Yunxia Chen, Prof., received a PhD degree in systems engineering from the School of Reliability and Systems Engineering, Beihang University, Beijing, China, in 2004. She is currently a Professor and the Vice Dean with the School of Reliability and Systems Engineering, Beihang University. Her main research interests include reliability design based on physics of failure and data, prognostics and health management, and experimental technology.



Dan Xu Assoc. Prof. received BS and PhD degrees from the School of Mechanical Engineering and Automation, Beihang University, Beijing, China, in 2003 and 2009, respectively. She is currently an Associate Professor with the School of Reliability and Systems Engineering, Beihang University. Her primary research interests include reliability design, assessment based on physics of failure, and degradation modeling.

Cong Wang received a BS degree from Beihang University, China, in 2020. He is currently pursuing a PhD degree with the School of Reliability and Systems Engineering at Beihang University. His current research interests are system reliability modeling and life prediction.

Weiyi Xiang received a BS degree from Beihang University, China, in 2024. He is currently pursuing a PhD degree with the School of Reliability and Systems Engineering at Beihang University. His research interests are centered on interpretability of machine learning models and its applications in fault diagnosis.

Xiaoqi Xiao received a BS degree from China University of Geosciences, China, in 2019, and the M.S. degree from Beihang University, in 2022. He is currently pursuing a PhD degree with the School of Reliability and Systems Engineering at Beihang University. His research interests include reliability assessment based on deep learning, degradation modeling.

Jing Lin, Prof., received BS, MS and PhD degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 1993, 1996 and 1999, respectively. From 2009 to 2018, he was a Professor with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. He is currently the Dean of the School of Reliability and Systems Engineering, Beihang University, Beijing, China. He is also the Changjiang Distinguished Professor with the Ministry of Education of China, Beijing, China. His current research field includes machinery condition monitoring, fault diagnosis and prognosis, vibration analysis, and nonstationary signal processing. Dr. Lin was the recipient of the Distinguished Young Scholar Funding from the National Natural Science Fund in 2011 and the State Natural Science Award in 2013.