Interdisciplinary Description of Complex Systems 24(6), 794-800, 2026

CRITICAL ARTIFICIAL INTELLIGENCE VALIDATION IN AUTONOMOUS VEHICLES – RECONSTRUCTION OF DECISION LOGIC AND RISK ANALYSIS

Laszlo Ady[1, *], Péter János Varga[2], György Schuster[2] and Dániel Tokody[1, 3]

[1]Óbuda University, Doctoral School on Safety and Security Sciences
Budapest, Hungary

[2]Óbuda University, Kandó Kálmán Faculty of Electrical Engineering
Budapest, Hungary

[3]NextTechnologies Ltd.
Maglód, Hungary

DOI: 10.7906/indecs.24.6.10
Regular article

Received: 17 June 2026.
Accepted: 25 July 2026.

ABSTRACT

The integration of Artificial Intelligence into Autonomous Vehicles presents a paradigm shift in transportation, promising enhanced safety and efficiency. However, the inherent complexity and "black-box" nature of many Artificial Intelligence models, particularly deep learning systems, pose unprecedented challenges for safety validation and certification. This article addresses the critical need for reconstructing the decision logic of Autonomous Vehicle Artificial Intelligence systems to enable rigorous quantitative risk analysis. We review the current landscape of Autonomous Vehicle incidents and regulatory standards, highlighting the gap between Artificial Intelligence capabilities and established safety integrity requirements. The core contribution of this article is the introduction of a novel, comprehensive risk metric R_{AI}, and a corresponding validation criterion, the Risk Balance Rate. These formulae integrate logic reconstruction, environmental context, and system criticality into a quantifiable framework for safety assessment. We further provide a qualitative comparison of different Artificial Intelligence model types – Traditional Machine Learning, Deep Learning, and Bayesian Networks – against the requirements for logic transparency, reliability, and risk. Our findings suggest that a hybrid approach, combining the strengths of different model paradigms with a robust reconstruction methodology, is essential for achieving the levels of validation required for public trust and regulatory approval. This article concludes that the proposed quantitative framework is a vital step toward standards-compliant, safe, and justifiable deployment of autonomous driving technology.

KEY WORDS

autonomous vehicles, artificial intelligence, safety validation, risk analysis, functional safety

CLASSIFICATION

APA: 4120

JEL: O33

*Corresponding author, η: adylaszlo@nexttechnologies.hu; -;
József körút 6., H – 1088 Budapest, Hungary

INTRODUCTION

The advent of autonomous vehicles (AVs) represents one of the most significant technological and societal shifts since the introduction of the automobile itself. By replacing human drivers with artificial intelligence (AI), these systems promise to drastically reduce traffic accidents caused by human error, increase mobility for the elderly and disabled, and optimize traffic flow for reduced congestion and emissions. At the heart of this promise lies the AI decision-making system, a complex software entity responsible for perceiving the environment, predicting the behavior of other road users, and planning a safe and efficient path.

However, the very technology that enables this leap in capability also introduces a fundamental challenge in safety engineering: how do we validate the decision-making of an AI system that is inherently non-deterministic and often functions as an opaque "black box"? Unlike traditional software, where logic is explicitly programmed and can be traced line-by-line, the behavior of a deep neural network is emergent and learned from vast datasets, making it unpredictable in novel or edge-case scenarios. The consequences of a validation failure are severe, potentially resulting in immediate collisions, loss of life, and a catastrophic erosion of public trust in the technology [1].

Recent history underscores the urgency of this challenge. High-profile incidents, including fatal crashes involving automated driving systems, have been linked to software logic failures and an inability of the AI to correctly interpret and react to unusual situations [2]. Regulatory bodies are beginning to respond, mandating the reporting of crashes involving driver-assistance systems, but a comprehensive framework for pre-deployment validation remains nascent [3]. Recent analyses have further quantified these risks, highlighting the specific safety vulnerabilities introduced by AI-driven solutions in road vehicles [4].

This article addresses this critical gap by proposing a methodology for reconstructing AI decision logic to enable quantitative risk analysis. The primary goal is to move beyond qualitative assessments and develop a framework that can objectively measure the safety of an AV's AI core. The article is structured as follows: Section 2 provides the necessary background on AV architectures and relevant safety standards. Section 3 introduces our core contribution: a novel mathematical framework for AI risk analysis, including the R_{AI} risk metric and the Risk Balance Rate (RBR) validation criterion. Section 4 offers a comparative analysis of different AI model types within this framework, incorporating insights from recent advances in Bayesian approaches and explainable AI. Finally, Section 5 concludes with a discussion of the implications for AV certification and future research directions.

BACKGROUND AND STANDARDS

Validating an autonomous vehicle is a multi-layered problem that extends far beyond the performance of its individual components. It requires a holistic view of the system's interaction with a complex, dynamic world.

In the context of autonomous vehicles, an AV operates in a continuous control loop. It must perceive its surroundings using sensors (cameras, LiDAR, radar), fuse this data into a coherent model of the world, predict how that world will evolve (e.g., will a pedestrian step into the crosswalk?), plan a trajectory, and finally execute that plan through vehicle controls [5]. The AI is the cognitive engine driving this loop, particularly in the perception, prediction, and planning stages.

The operational environment is characterized by:

- Dynamic Urban Environments – The system must contend with a chaotic mix of static and moving objects, varying lighting and weather conditions, and unpredictable human behavior.
- Mixed Traffic – The AI must safely interact with other autonomous systems, human-driven vehicles, pedestrians, cyclists, and scooters, each with its own level of predictability.
- Real-Time Constraints – Decisions must be made and acted upon in milliseconds. A delay in perception or planning can be as dangerous as an incorrect decision.

The ongoing problem of safety validation. Traditional automotive safety validation relies heavily on extensive real-world testing and milage accumulation. However, this approach is insufficient for AI-based systems. It is statistically impossible to test for every conceivable edge case, and the system's behavior can change in ways that are not captured by simple pass/fail criteria.

The core challenges are summarised as follows:

- Logic Reconstruction – For safety assurance, we must understand why an AI made a particular decision. This is difficult for "black-box" models like deep neural networks, where the decision logic is distributed across millions of weights and activations.
- Robustness Testing – The AI must be resilient to small, often imperceptible, changes in input data (adversarial examples) that could cause it to make catastrophic errors, such as misclassifying a stop sign [6].
- Formal Verification – Proving that an AI system will always behave within a set of safe boundaries is an extremely difficult, and in some cases undecidable, problem.

Relevant industry standards can be mentioned in this context. In the automotive industry has a well-established framework for functional safety, primarily defined by ISO 26262 [7]. This standard introduces the concept of Safety Integrity Levels (SIL), or more specifically for automotive, Automotive Safety Integrity Levels (ASIL), which classify the necessary risk reduction for a system based on the severity, exposure, and controllability of hazardous events, Table 1. However, as Salay et al. [8] analyze in detail, applying ISO 26262 to machine learning components presents significant challenges, particularly concerning the verification of software whose functionality is learned from data rather than specified.

Table 1. Safety integrity level target failure rates.

SIL	Probability of Failure per Hour Target
1	10^{-5}
2	10^{-6}
3	10^{-7}
4	10^{-8}

While ISO 26262 provides a robust framework for hardware and software developed in a traditional manner, it does not fully address the unique challenges of AI-based systems, which are increasingly being covered by emerging standards like ISO 21448 Safety Of The Intended Functionality (SOTIF) [9]. SOTIF focuses on ensuring the safety of a function even in the absence of a system failure, specifically addressing performance limitations and inadequacies in the design or implementation, which is highly relevant for AI perception and decision errors. Recent comprehensive studies have mapped the landscape of these safety risks, confirming the critical interplay between ISO 26262, SOTIF, and the forthcoming AI-specific guidelines such as ISO PAS 8800 [4].

The goal of logic reconstruction, therefore, is to bridge the gap between these standards. By systematically mapping AI inputs to outputs, we can begin to build the evidence necessary to claim that a system meets a target SIL or is free from unreasonable risk under its intended functionality.

A FRAMEWORK FOR QUANTITATIVE RISK ANALYSIS

To move towards objective, standards-compliant validation, we propose a quantitative framework that decomposes AI risk into measurable components. This framework is built upon two core contributions: a general risk metric for AI in AVs, and a validation criterion for deployment decisions.

We have defined a General AI Risk Metric (R_{AI}). This metric quantifies the general risk posed by an AI decision-making system in an autonomous vehicle.

$$R_{AI} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\phi(x_i, y_i)}_{\text{Data Risk}} \times \underbrace{\Gamma(M, A)}_{\text{Logic Reconstruction}} \times \underbrace{\Psi(E)}_{\text{Traffic Env.}} \times \underbrace{e^{\lambda \cdot \text{SIL}}}_{\text{Criticality}} \quad (1)$$

This equation models risk as a product of four key factors, aggregated over a set of n driving scenarios or decisions.

$\phi(x_i, y_i)$ is Validation Factor. This term represents the "surprise" or distributional shift between the training/validation data and the real-world operational data. For a given scenario i, x_i is the input (e.g., sensor data) and y_i is the system's output. The function $\phi(\cdot, \cdot)$ measures how far this input-output pair lies from the distribution of data the system was validated on. A high value indicates the system is operating in a novel or poorly-understood regime, increasing uncertainty and therefore risk. This directly addresses the SOTIF concern of performance limitations.

Γ(M, A) is Logic Reconstruction Factor. This is the core of our proposal. Γ quantifies the complexity of reconstructing the decision logic for a given model M and reconstruction methodology A. A model with high transparency (e.g., a simple decision tree) would have a low Γ value, as its logic is easily understood. A "black-box" deep neural network, where reconstruction requires complex techniques like feature visualization or surrogate models, would have a high Γ value. The factor Γ thus acts as a multiplier, increasing the assessed risk for models whose decision-making processes are opaque and difficult to verify. The emerging field of eXplainable AI provides tools to potentially reduce this factor, as explored in recent work on world models for driver assistance [10].

Ψ(E) is the Environmental Pressure Factor. This term accounts for the complexity and riskiness of the traffic environment E. It could be a function of variables such as traffic density, number of road users, weather conditions, and intersection complexity. An AI operating on a sparsely trafficked highway experiences lower environmental pressure (Ψ) than one navigating a dense, chaotic urban intersection. This factor contextualizes the decision within its operational design domain.

Finally, $e^{\lambda \cdot \text{SIL}}$ in (1) is the System Criticality Factor. This exponential term incorporates the required safety integrity level of the function being performed. The SIL value, derived from a hazard analysis and risk assessment (e.g., ASIL determination), represents the inherent criticality of the task. The parameter λ is a scaling factor that calibrates the sensitivity of the risk metric to the SIL. A higher SIL (e.g., SIL = 3 for an emergency braking function) results in a much larger multiplier, ensuring that functions which require a higher degree of integrity are subjected to a much stricter risk threshold.

THE RISK BALANCE RATE VALIDATION CRITERION

The R_{AI} metric, by itself, provides a measure of risk. To make a deployment decision, this risk must be weighed against the expected safety benefit of introducing the automation. We propose the RBR as this decision criterion:

$$RBR = \frac{R_{\text{ARRP}}\ \text{(Risk Reduction)}}{R_{\text{ARAR}}\ \text{(AI-induced Risk)}} \tag{2}$$

where:

- R_{ARRP} (Risk Reduction – Automation Risk Reduction Potential) represents the quantified safety benefit of the AI system. It is the theoretical reduction in risk compared to a baseline, typically human driving. This could be calculated based on statistics of human-error-related accidents that the automation is designed to prevent.
- R_{ARAR} (AI Induced Risk) is the risk introduced by the AI system itself. For a given deployment scenario, we can use the R_{AI} metric as a primary input to R_{ARAR}, potentially combined with other system-level failure rates.

The RBR provides a simple, interpretable ratio for decision-making. The RBR greater than 1 indicates that the risk reduction potential outweighs the induced risk, justifying deployment, Table 2. Conversely, RBR less than 1 suggests the system is not yet safe enough.

Table 2. AV deployment decision based on RBR.

RBR Range	Validation Status	Operational Outcome
High (RBR > 1,5)	Validated Decision Logic	Full Autonomous Mode
Medium (1,0 ≤ RBR ≤ 1,5)	Conditional Reliability	Human-in-the-loop (supervision required)
Marginal (0,8 ≤ RBR < 1,0)	Insufficient Safety	Restricted / Testing / Safety Driver / Shadow Mode
Low (RBR < 0,8)	Unreliable Logic	Emergency Braking Only / Fallback Mode

COMPARATIVE ANALYSIS OF AI MODELS FOR AV LOGIC

The choice of AI model architecture has a profound impact on the terms in our risk equations, particularly the logic reconstruction factor Γ. Different models offer different trade-offs between performance, transparency, and inherent risk, Table 3. Recent advances, particularly in Bayesian methods, are reshaping this landscape by offering robust ways to handle uncertainty.

Table 3. Qualitative comparison of models in AV context.

Model Type	Logic Transparency	AV Reliability	Inherent Risk Score
Traditional ML (e.g., SVM, Decision Trees)	High	Medium	Medium
Deep Learning (e.g., CNNs, Transformers)	Low	High	High
Bayesian Networks	Medium	High	Low

In the case of traditional machine learning models, such as support vector machines and decision trees, models like decision trees offer explicit, human-readable decision paths. This makes logic reconstruction relatively straightforward, as the factors influencing a decision can be traced directly, Table 3.

While robust in well-defined scenarios, traditional ML models often struggle with the high-dimensional, unstructured data (like camera images) that is essential for complex scene understanding. Their performance may plateau compared to more modern approaches, Table 3.

The high transparency helps lower the risk score by making the system easier to validate and debug. However, their potential performance limitations in complex environments can be a source of risk, Table 3.

Deep learning models (e.g. convolutional neural networks and transformers) are the quintessential 'black boxes'. Understanding why a network classified an image a certain way or predicted a specific trajectory is a major research challenge. This results in a very high Γ factor. Deep learning has enabled the leap in AV capabilities. Its performance in perception (object detection, lane detection) and prediction tasks far exceeds traditional methods, leading to high operational reliability in a statistical sense.

The combination of high performance (which encourages their use) and extremely low transparency leads to a high inherent risk score. The potential for silent, unpredictable failures in edge cases is a major safety concern.

Bayesian networks model probabilistic relationships between variables in a graphical structure. This structure is more transparent than a neural network, as the dependencies can be visualized and understood. However, the exact probabilistic calculations can become complex.

Bayesian networks are excellent at handling uncertainty and combining different sources of evidence (sensor fusion). They can provide robust predictions even with incomplete or noisy data, which is a significant advantage in real-world driving.

The probabilistic nature of Bayesian networks allows for a more nuanced representation of risk and confidence. Their outputs are not just a single decision but a distribution of possibilities, which can be used to implement safer, more cautious behavior. This inherent handling of uncertainty contributes to a lower overall risk score in our framework.

This comparison suggests that the future of safe AV AI may lie in hybrid architectures. For instance, a deep learning-based perception module could feed its outputs, along with uncertainty estimates, into a Bayesian network for decision-making and planning. This would leverage the perceptual power of deep learning while mitigating its transparency and risk shortcomings with the probabilistic reasoning and relative transparency of a Bayesian system.

CONCLUSION

The safe and widespread deployment of autonomous vehicles hinges on our ability to validate the AI that controls them. The "black-box" nature of state-of-the-art AI models presents a fundamental challenge to traditional safety engineering paradigms. This article has argued that a paradigm shift towards quantitative, reconstruction-based validation is essential.

Our contribution is a novel, mathematically grounded framework that moves the conversation from qualitative assurance to quantitative analysis. The proposed R_{AI} risk metric provides a structured way to decompose AI risk into its constituent parts: validation uncertainty, logic complexity, environmental context, and system criticality. The companion RBR offers a clear, actionable criterion for making deployment decisions, ensuring that a system's safety benefits demonstrably outweigh its induced risks.

Our comparative analysis underscores that model choice is a critical safety decision. While deep learning offers unparalleled performance, its opacity leads to a high inherent risk score, making rigorous logic reconstruction not just beneficial, but mandatory. Approaches like Bayesian networks, or hybrid models that combine the strengths of different paradigms, offer a promising path toward more verifiable and trustworthy AI systems.

Logic reconstruction is not merely an academic exercise; it is an essential prerequisite for AV certification. By making the invisible visible, we can build the safety case necessary to satisfy regulators and, most importantly, earn the public's trust. The formula for the future of transportation is simple: Safety First. And safety first demands a commitment to understanding the very intelligence that will be entrusted with our lives on the road.

ACKNOWLEDGMENTS

Project no. 2023-2.1.2-KDP-2023-00009 implemented with support from the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2023 funding scheme.

REFERENCES

[1] Kalra, N. and Paddock, S.M.: *Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?* https://www.rand.org/pubs/research_reports/RR1478.html, accessed 30th January 2026,

[2] National Transportation Safety Board: *Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston.* https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf, accessed 30th January 2026,

[3] National Highway Traffic Safety Administration: *Standing General Order on Crash Reporting for Level 2 ADAS and Level 3-5 Automated Driving Systems.* https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting, accessed 30th January 2026,

[4] Mirzarazi, F.; Danishvar, S. and Mousavi, A.: *The Safety Risks of AI-Driven Solutions in Autonomous Road Vehicles*. World Electric Vehicle Journal **15**(10), No. 438, 2024, http://dx.doi.org/10.3390/wevj15100438,

[5] Paden, B.; Čáp, M.; Yong, S.Z.; Yershov, D. and Frazzoli, E.: *A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles*. IEEE Transactions on Intelligent Vehicles **1**(1), 33-55, 2016, http://dx.doi.org/10.1109/TIV.2016.2578706,

[6] Goodfellow, I.J.; Shlens, J. and Szegedy, C.: *Explaining and harnessing adversarial examples.* preprint arXiv:1412.6572v3 [stat.ML], https://doi.org/10.48550/arXiv.1412.6572,

[7] International Organization for Standardization: *ISO 26262-1:2018 Road vehicles – Functional safety*. 2nd edition. https://www.iso.org/standard/68383.html, accessed 30th January 2026,

[8] Salay, R.; Queiroz, R. and Czarnecki, K.: *An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software*. SAE Technical Paper 2018-01-1075, 2018, http://dx.doi.org/10.4271/2018-01-1075,

[9] International Organization for Standardization: *ISO 21448:2022 Road vehicles – Safety of the intended functionality*. https://www.iso.org/standard/77490.html, accessed 30th January 2026,

[10] Roshdi, M., et al.: *On the Road to Clarity: Exploring Explainable AI for World Models in a Driver Assistance System*. preprint arXiv:2404.17350v1 [cs.LG], http://dx.doi.org/10.48550/arXiv.2404.17350.