Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Revolutionary Approaches to Functional Safety in AI-Enabled Embedded Systems

Narendra Babu Atmakuri
Independent Researcher, USA
Corresponding Author: Narendra Babu Atmakuri, **E-mail**: natmakuri9@gmail.com

| ABSTRACT

The artificial intelligence in safety-critical embedded systems has required a redesign of functional safety concepts. This article discusses revolutionary approaches to the problem of the probabilistic nature of AI that introduces unprecedented challenges in application areas, where classical deterministic standards of safety are inapplicable. It examines novel architectures of safety, such as independent oversight systems and redundancy inference systems that offer life-saving safety measures to the AI components. The specialized verification techniques, which adapt formal methods to neural networks, and runtime monitoring methods, which can identify possible security breaches during operation, are reviewed. The article examines how to extend an existing set of standards, such as ISO 2626,2 to be compatible with the specific use of AI and how to create safety-aware training processes, which introduce safety constraints into the training process. The most promising directions to achieve certification are illustrated through explainable AI methods, which provide visibility into safety validation and hybrid systems, considering rule-based systems and AI capabilities in combination. Applying case studies in the automotive, aerospace, and medical sectors, this article describes how the use of such complementary methods can allow safe application of AI in highly regulated sectors to ensure that strict safety requirements are met, even where intrinsic limitations on verification are acknowledged.

| KEYWORDS

Safety-critical AI, Runtime monitoring, Formal verification, Hybrid safety architectures, Explainable AI, Safety-aware training

| ARTICLE INFORMATION

ACCEPTED: 01 July 2025 **PUBLISHED:** 31 July 2025 **DOI:** 10.32996/jcsts.2025.7.8.29

1. Introduction

The advent of artificial intelligence in safety-critical embedded systems has seriously shaken traditional functional safety practices. Matters get more serious with critical applications in the automotive, aerospace, healthcare, and industrial fields that require AI technologies to a greater extent, the engineers have to contend with making sure that the operations are safe, regardless of the probabilistic nature of machine learning models. Such drastic change necessitates the rethinking of the principles of safety assurance.

Neural networks introduce profound unpredictability—minor input variations sometimes yield wildly different outputs, creating formidable challenges for safety-critical implementations. Traditional frameworks like ISO 26262 and IEC 61508 were crafted for deterministic systems with predictable execution paths, whereas AI systems function through statistical inference using billions of parameters. Bridging this fundamental disconnect requires innovative safety assurance approaches tailored to machine learning's unique characteristics [1].

Recent investigations have centered on developing specialized safety frameworks for AI-enabled embedded systems. Discussions at the Workshop on Reliable and Interpretable Artificial Intelligence emphasized that safety-critical neural networks must demonstrate qualities beyond mere accuracy, including adversarial input robustness, distribution shift resilience, decision interpretability, and uncertainty quantification. Conventional verification techniques inadequately address these requirements, leaving significant safety assurance gaps [1].

Runtime monitoring frameworks stand among the most promising AI safety approaches. These mechanisms continuously assess AI components during operation, identifying potential safety violations and triggering appropriate fallback mechanisms when necessary. For automated driving applications, such frameworks establish explicit safety conditions required for normal operation, alongside clear fallback strategies for violation scenarios [2].

Hybrid architectures blending traditional rule-based systems with AI components offer another groundbreaking approach. These designs maintain separation between AI and safety-critical functions, implementing dedicated monitoring systems overseeing AI components. This strategy tackles the fundamental challenge of comprehensively verifying complex neural networks by constraining AI authority while maintaining verified fallback mechanisms [2].

Formal verification techniques adapted for neural networks deliver mathematical guarantees regarding behavior within defined operational boundaries. Though complete verification remains elusive for large networks, substantial progress has emerged in verifying critical safety properties within constrained domains. These techniques complement runtime monitoring and hybrid architectures to build comprehensive safety cases [1].

This article explores these pioneering research directions and practical applications across multiple domains. From safety-aware training methodologies to sophisticated runtime monitors, these approaches mark a fundamental evolution in functional safety thinking, paving the way for safe deployment of increasingly autonomous and capable AI systems in safety-critical applications.

2. Beyond Deterministic Safety Analysis

Traditional functional safety standards, including ISO 26262 (automotive), IEC 61508 (industrial), and DO-178C (aerospace), were developed for deterministic systems with predictable failure modes. AI-enabled systems introduce non-deterministic behaviors necessitating fundamentally new safety approaches. Conventional standards rely on systematic hazard analysis, fault tree assessments, and exhaustive testing—methodologies proving fundamentally inadequate when applied to systems incorporating neural networks and machine learning techniques operating probabilistically rather than deterministically [3].

The limitations of traditional safety approaches become glaringly apparent when examining the statistical nature of AI decision-making. Examinations into AI safety assurance reveal substantial challenges in applying conventional verification methods to neural networks. Traditional safety engineering aims to demonstrate that systems never enter specified unsafe states under any circumstances, yet modern deep learning systems lack the theoretical foundations enabling such guarantees. This creates a fundamental verification gap impossible to address through conventional testing methodologies alone, necessitating complementary approaches specifically designed for AI systems [3].

Recent investigations demonstrate that safety assurance for AI-enabled embedded systems demands fresh methodologies accommodating statistical performance guarantees rather than binary pass/fail criteria. Unlike traditional systems where component behavior undergoes exhaustive verification against requirements, AI components require evaluation through probabilistic performance across operational domains. This shift represents a fundamental reconceptualization of safety engineering, moving from guarantees toward statistical confidence bounds with explicitly quantified uncertainty. Emerging frameworks for establishing AI safety focus on bounding worst-case behaviors while implementing runtime monitoring systems capable of detecting and mitigating potential safety violations during operation [4].

AI model performance degradation over time due to distribution shifts presents another challenge, unaddressed by traditional safety standards. Research on AI safety for autonomous systems documents substantial performance declines in deployed models as operational conditions diverge from training data distributions. This "model aging" phenomenon necessitates continuous monitoring and adaptation strategies absent from conventional safety frameworks. Similarly, machine learning systems exhibit novel failure modes emerging from complex interactions between trained models and unexpected inputs—failure patterns defying anticipation during development despite rigorous testing [4].

These challenges in providing complete verification and validation coverage for AI-enabled systems have stimulated the development of multi-layered safety approaches combining design-time verification with runtime monitoring and graceful degradation strategies. Rather than attempting to prove system safety exclusively through pre-deployment verification, these approaches acknowledge fundamental limitations of AI verification while implementing continuous safety monitoring during operation. This represents a paradigm shift in safety thinking, moving from prevention-focused approaches toward detection and mitigation strategies, maintaining safety despite inherent uncertainty in AI components [3].

Challenge	Traditional Approach	AI-Enabled Approach	Safety Impact
Verification Method	Exhaustive Testing	Statistical Guarantees	Medium
Performance Assessment	Binary Pass/Fail	Probabilistic Performance	High
Long-term Reliability	Static Validation	Continuous Monitoring	Very High
Failure Mode Analysis	Predetermined Patterns	Dynamic Adaptation	High
Safety Assurance	Prevention-focused	Detection & Mitigation	High

Table 1: Paradigm Shift: From Deterministic to Probabilistic Safety Assurance [3, 4]

3. Innovative Safety Architectures

3.1 AI Safety Monitors

Among the most promising architectural patterns emerging involves implementing dedicated safety monitoring systems operating independently from primary AI inference engines. These monitors continuously evaluate operating conditions and outputs of AI components against predefined safety envelopes. This approach acknowledges fundamental verification limitations of complex AI systems while implementing runtime safeguards capable of detecting and mitigating potential safety violations during operation. Research on model-driven safety assurance for deep learning systems demonstrates that properly designed safety monitors provide a critical safety layer for AI-enabled systems, enabling deployment in highly regulated domains while maintaining strong safety guarantees [5].

Safety monitors draw inspiration from the Simplex architecture developed for traditional control systems, while extending this approach to address the unique challenges of AI components. Unlike conventional software, where safety properties undergo verification through exhaustive testing or formal methods, neural networks require continuous monitoring to ensure safe operation across all possible operational conditions. These monitoring systems operate independently from primary AI components, using separate sensing pathways and simplified algorithms, allowing more thorough verification. Model-driven approaches to safety monitoring provide systematic methods for designing and implementing these safeguards, ensuring detection of potentially dangerous conditions before harm occurs [5].

Research across multiple safety-critical domains has converged on several effective monitoring approaches addressing different aspects of AI safety. Uncertainty quantification mechanisms assess the confidence of AI predictions, allowing systems to detect situations where AI operates outside validated domains. Runtime verification techniques compare AI outputs against physics-based models encoding fundamental constraints about physical worlds, identifying predictions violating these constraints. Anomaly detection systems identify behavioral patterns deviating from expected norms, potentially indicating novel failure modes not anticipated during development. Input validation frameworks detect out-of-distribution scenarios where AI encounters inputs significantly different from training data, triggering appropriate fallback mechanisms before unsafe decisions occur [6].

These monitoring approaches see implementation across multiple industries, with particularly significant advances in autonomous vehicles and medical diagnostics. By establishing clear safety boundaries and implementing mechanisms for detecting boundary violations, safety monitors provide a critical protection layer, compensating for the inherent verification limitations of AI systems. Recent research on testing deep learning systems emphasizes the importance of these monitoring approaches within comprehensive safety frameworks, enabling deployment of AI capabilities in safety-critical domains while maintaining strong safety guarantees through continuous monitoring and fallback mechanisms [6].

3.2 Redundant Inference Architectures

Another revolutionary approach involves deploying multiple, diverse AI implementations running in parallel to perform identical inference tasks. This approach draws inspiration from traditional N-modular redundancy used in safety-critical hardware while adapting the concept to address the unique characteristics of AI systems. Unlike traditional redundancy, where identical components undergo replication to protect against random hardware failures, AI redundancy focuses on diversity, protecting against systematic errors in training data, model architecture, or implementation. Model-driven safety assurance frameworks

demonstrate that properly designed diverse redundant architectures significantly reduce the probability of common-mode AI failures compared to single-model implementations [5].

These systems employ architectural diversity through multiple complementary strategies addressing different aspects of AI reliability. Heterogeneous model architectures combine fundamentally different approaches to identical problems, such as convolutional neural networks alongside transformer models, ensuring architectural weaknesses in one approach are unlikely to be shared by others. Training diversity uses different datasets or training methodologies, creating models with complementary strengths and weaknesses. Implementation diversity leverages different frameworks or hardware accelerators, protecting against implementation-specific vulnerabilities. Sophisticated voting mechanisms resolve conflicts between model outputs, implementing consensus algorithms identifying and rejecting outlier predictions while maintaining high overall system performance [6].

The effectiveness of redundant inference architectures has been demonstrated across multiple safety-critical domains. A recent automotive ADAS implementation demonstrated that triple-redundant neural networks with diverse architectures achieved safety integrity levels comparable to traditional hardware redundancy approaches while maintaining performance advantages of AI-based perception. Similar approaches see implementation in medical diagnostic systems, where diverse model ensembles show significant improvements in error detection compared to single-model implementations. Research on testing deep learning systems identifies these redundant architectures as key strategies addressing unique verification challenges of AI components, providing effective pathways to deploying AI capabilities in safety-critical applications while meeting stringent safety requirements [6].

Combining diverse redundant architectures with independent safety monitors creates particularly robust safety frameworks for AI-enabled systems. By implementing multiple complementary safety layers, these approaches address fundamental verification challenges of AI systems while enabling deployment in domains with stringent safety requirements. Model-driven approaches provide systematic methods for designing and implementing these safety architectures, ensuring strong safety guarantees despite the inherent uncertainty of AI components. This multi-layered approach represents a revolutionary advance in functional safety thinking, moving beyond traditional deterministic approaches to embrace the statistical nature of AI while maintaining strong safety guarantees [5].

Architecture Component	Implementation Approach	Safety Benefit	Application Domain
Safety Monitors	Independent Verification	Real-time Protection	Automotive, Medical
Uncertainty Quantification	Confidence Assessment	Out-of-Domain Detection	All Domains
Physics-Based Verification	Constraint Validation	Reality Alignment	Robotics
Redundant Inference	Diverse Model Architectures	Common-Mode Failure Protection	Aerospace, Automotive
Heterogeneous Models	Multi-paradigm Approaches	Architectural Diversity	Safety-Critical Systems

Table 2: Multi-layered Safety Frameworks for AI Components [5, 6]

4. Verification Techniques for Neural Network Behavior

Formal verification of neural networks constitutes an active research area marked by substantial recent breakthroughs. The central issue in neural network verification is due to the complex non-linear nature and the sheer expansive degree of input-error-output mappings. Formal verification techniques based on complete testing or on exhaustive methods that analyse the code (static analysis) are computationally challenging to apply to current deep learning models with millions of parameters or billions of parameters. This reality has driven the development of specialized verification techniques crafted specifically for neural networks, adapting formal methods concepts while tackling machine learning's distinctive challenges [7].

4.1 Formal Methods Adaptation

Specialized verification tools delivering mathematical guarantees about neural network behavior within specified operational domains mark significant progress in AI safety. These instruments enable rigorous verification of critical properties impossible to establish through testing alone. By furnishing mathematical guarantees rather than statistical confidence derived through testing,

such approaches facilitate neural network certification for safety-critical applications across highly regulated sectors, including automotive and aerospace systems [7].

Abstract interpretation techniques stand among the most promising neural network verification approaches. These methods compute over-approximations of possible outputs for given input ranges, establishing behavioral bounds without demanding exhaustive analysis across all potential inputs. Investigations into formal verification of neural networks demonstrate abstract interpretation's effectiveness for verifying robustness properties of convolutional neural networks, establishing formal guarantees that certain adversarial input classes cannot trigger misclassification. These techniques have successfully verified perception systems for autonomous vehicles, providing mathematical assurances that objects receive correct classification across varied environmental conditions [8].

SMT (Satisfiability Modulo Theories) solvers adapted for neural network verification allow formulating verification problems as constraint satisfaction problems amenable to efficient solution. Research on decision procedures for automated analysis has extended traditional SMT solving approaches to handle neural networks' distinctive challenges, including non-linear activation functions and high-dimensional input spaces. These approaches have successfully verified safety properties for neural network controllers in autonomous systems, establishing formal guarantees that control actions maintain safety constraints throughout operational domains. The capacity to verify complex properties through SMT solving marks a significant advancement in neural network verification, enabling rigorous safety analysis unattainable through testing alone [7].

Reachability analysis techniques prove safety properties of closed-loop AI control systems, addressing challenges in verifying systems where neural networks interact with physical environments. These approaches compute reachable state sets from given initial conditions, enabling verification that safety constraints remain preserved throughout system operation. Research on verifying deep neural networks demonstrates reachability analysis effectiveness for verifying neural network controllers in robotic systems, establishing formal guarantees that robots maintain safe distances from obstacles despite sensing and control uncertainties. These techniques create pathways toward verifying complex AI systems interacting with physical environments, a critical requirement for autonomous systems in safety-critical domains [8].

4.2 Runtime Monitoring Systems

Beyond design-time verification, runtime monitoring systems constitute critical safety layers for deployed AI systems. These mechanisms continuously evaluate AI components during operation, detecting potential safety violations and activating appropriate mitigation strategies. Runtime monitoring compensates for inherent design-time verification limitations, providing supplementary safety layers that detect and mitigate potential safety violations during operation. This approach acknowledges fundamental difficulties in exhaustively verifying neural networks before deployment, and implementing continuous safety evaluation during operation [7].

Online distribution shift detection algorithms identify situations where models operate beyond validated domains, addressing critical safety concerns for deployed AI systems. Research on satisfiability modulo theories demonstrates statistical monitoring approaches' effectiveness for real-time distribution shift detection, enabling systems to activate fallback mechanisms before unsafe decisions materialize. These techniques continuously compare operational input statistical properties against training data properties, identifying significant deviations potentially indicating reduced model reliability. By detecting distribution shifts before safety violations occur, these approaches enable safe AI system deployment in dynamic environments where operating conditions evolve [7].

Adversarial input detection mechanisms recognize potential malicious manipulations or naturally occurring inputs that potentially cause unexpected model behavior. Research on verifying deep neural networks has developed effective detection algorithms identifying adversarial inputs with high accuracy, allowing systems to reject potentially dangerous inputs before unsafe decisions result. These approaches combine multiple detection strategies, including input pattern statistical analysis, comparison against known adversarial signatures, and model uncertainty evaluation for specific inputs. By detecting and rejecting adversarial inputs, these systems protect against malicious attacks, and naturally occurring inputs potentially trigger unsafe behaviors [8].

Graceful degradation pathways activate when AI components cannot operate safely, maintaining system safety through fallback mechanisms with reduced functionality. Research on reluplex verification algorithms demonstrates effective degradation strategies for autonomous vehicles, enabling continuous safe operation despite temporary AI perception system failures. These approaches implement multi-level fallback mechanisms progressively reducing system capability while preserving safety, from reduced speed operation to complete stops at safe locations. Through structured degradation pathway implementation, these systems ensure safety preservation even when AI components cannot operate as intended, a critical requirement for deploying AI in safety-critical applications [8].

Combining formal verification techniques with runtime monitoring systems creates comprehensive approaches to neural network verification, addressing both design-time and runtime safety concerns. By establishing formal guarantees about system behavior where possible while implementing continuous monitoring for conditions potentially violating these guarantees, this approach enables safe AI system deployment in highly regulated domains. This marks a significant advancement in functional safety for AI-enabled systems, adapting traditional verification concepts to address machine learning's unique challenges while maintaining strong safety guarantees [7].

Verification Technique	Methodology	Application Area	Verification Strength
Abstract Interpretation	Output Approximation	Perception Systems	Medium
SMT Solvers	Constraint Satisfaction	Control Systems	High
Reachability Analysis	State Space Exploration	Closed-loop Systems	High
Runtime Monitoring	Continuous Evaluation	All AI Systems	Medium
Distribution Shift Detection	Statistical Comparison	Deployed Systems	Medium
Graceful Degradation	Fallback Mechanisms	Autonomous Systems	Very High

Table 3: Complementary Verification Strategies for Neural Networks [7, 8]

5. Extending Functional Safety Standards

The research community has achieved substantial progress in extending traditional safety standards to accommodate AI components. The standards of existing functional safety, such as ISO26262 (vehicle), IEC61508 (industrial), and DO-178C (aerospace), were designed to apply to conventional software in a deterministic environment with known failure modes. These standards are based on organized laboratory hypotheses of hazard analysis, assessment, and verification processes, which are dependent on the assumptions that software actions are deterministic and thus can be covered by exhaustion testing. AI components, particularly deep learning-based systems, violate these fundamental assumptions, creating significant certification challenges under existing standards [9].

Integrating AI into safety-critical systems has necessitated a fundamental rethinking of functional safety approaches. Rather than completely abandoning established standards, researchers have focused on extending these frameworks to address unique AI component characteristics while maintaining compatibility with existing safety processes. This evolutionary approach enables certification of AI-enabled systems within established regulatory frameworks, facilitating adoption of AI capabilities in highly regulated domains while ensuring safety [10].

5.1 ISO 26262 Extensions

For automotive applications, researchers have proposed concrete ISO 26262 extensions addressing unique AI component characteristics. ISO 26262, the international standard for functional safety in road vehicles, provides comprehensive frameworks ensuring the safety of electrical and electronic systems in automobiles. However, applying this standard to AI components presents significant challenges due to machine learning systems' non-deterministic nature and difficulties in providing complete verification evidence through traditional means [9].

Adaptation of ASIL (Automotive Safety Integrity Level) determination methods for machine learning components represents one significant ISO 26262 extension. Traditional ASIL determination assesses severity, exposure, and controllability of potential hazards, resulting in integrity requirements ranging from ASIL A (lowest) to ASIL D (highest). For AI components, researchers propose modified determination methods accounting for machine learning decisions' probabilistic nature and potential unexpected behaviors in novel situations. Analysis of ISO 26262 in machine learning safety contexts identifies specific gaps in the standard's hazard analysis and risk assessment approach when applied to neural networks. Researchers have developed extensions incorporating factors such as model uncertainty, operational domain coverage, and distribution shift robustness into ASIL determination processes, enabling appropriate safety requirements for AI components [9].

Developing safety cases incorporating evidence from statistical performance guarantees represents another important ISO 26262 extension. Traditional safety cases heavily rely on deterministic evidence such as formal verification results and exhaustive test coverage metrics. For AI components, researchers have developed complementary approaches incorporating statistical performance guarantees, uncertainty quantification, and runtime monitoring data into comprehensive safety arguments. Research on safety assurance for machine learning in automotive software identifies specific challenges creating safety cases for machine learning components, proposing structured approaches integrating diverse evidence sources into compelling safety arguments satisfying regulatory requirements while acknowledging inherent AI verification limitations [9].

Updated Failure Mode and Effects Analysis (FMEA) methodologies for neural networks address challenges in identifying and mitigating potential failure modes in AI components. Traditional FMEA approaches struggle to capture complex neural network failure patterns, which may exhibit unexpected behaviors due to distribution shifts, adversarial inputs, or complex model interactions. Research on ethical and social implications of AI highlights systematic neural network failure analysis, identifying AI-specific failure modes requiring addressing in safety-critical applications. These approaches enable systematic analysis and mitigation of potential failure modes in AI components, a critical requirement for certification under ISO 26262 [10].

5.2 Safety-Aware AI Training Methodologies

Novel training approaches explicitly incorporating safety considerations during model development represent fundamental shifts in machine learning practices for safety-critical applications. Traditional machine learning optimization primarily focuses on performance metrics such as accuracy or precision, potentially sacrificing safety-critical properties in pursuit of overall performance. Safety-aware training methodologies explicitly incorporate safety constraints and objectives into training processes, creating models balancing performance with safety considerations [9].

Safety-constrained optimization objectives penalizing potentially unsafe decisions represent direct approaches to safety-aware training. Traditional neural network training typically employs loss functions focused on performance metrics such as classification accuracy or regression error. Safety-constrained optimization extends these loss functions with additional terms penalizing outputs that potentially lead to unsafe system behavior. Research on using machine learning safely in automotive software identifies specific approaches implementing safety-constrained optimization, including specialized loss functions prioritizing safety-critical performance. These approaches enable the development of models prioritizing safety-critical performance while maintaining overall accuracy [9].

Robust training techniques ensure performance across worst-case operational scenarios, addressing traditional neural networks' vulnerability to adversarial inputs and distribution shifts. These approaches incorporate adversarial examples, synthetic edge cases, and domain randomization into training processes, creating models that maintain performance across wide operational condition ranges. Research on ethical and social implications of AI emphasizes robust training for safety-critical applications, identifying specific techniques improving model resilience against distribution shifts and adversarial inputs. These approaches prove particularly valuable for autonomous systems operating safely across diverse and potentially novel environments [10].

Safety curriculum learning prioritizes training on safety-critical edge cases, ensuring models develop appropriate responses to rare yet potentially dangerous situations. Traditional curriculum learning gradually increases task difficulty during training to improve overall learning efficiency. Safety curriculum learning extends this concept, prioritizing exposure to safety-critical scenarios, ensuring models receive sufficient training on rare but important cases potentially underrepresented in naturally distributed training data. Analysis of machine learning safety for automotive applications identifies specific approaches implementing safety curriculum learning, demonstrating effectiveness in improving performance on safety-critical edge cases [9].

Verification-guided training incorporates formal verification feedback loops, steering models toward formally verifiable behavior. This approach represents tight integration between verification and training, where verification results inform subsequent training iterations. Research on ethical and social implications of AI identifies verification-guided training as a promising approach for developing models with verifiable safety properties. This approach creates models more amenable to formal verification, enabling stronger safety guarantees for critical properties. By bridging gaps between training and verification, this approach addresses fundamental challenges certifying AI components for safety-critical applications [10].

Combining extended safety standards with safety-aware training methodologies creates comprehensive frameworks for developing and certifying AI components in safety-critical systems. By extending established standards while developing specialized training approaches for safety-critical applications, researchers have created pathways safely integrating AI capabilities into highly regulated domains. These approaches acknowledge unique AI component characteristics while leveraging established safety engineering principles, enabling certification of AI-enabled systems under existing regulatory frameworks [10].

Standard Extension	Traditional Approach	AI Extension	Certification Impact
ASIL Determination	Deterministic Assessment	Uncertainty Incorporation	High
Safety Case Construction	Exhaustive Evidence	Statistical Guarantees	Very High
Failure Analysis	Predefined Patterns	AI-Specific Failure Modes	High
Safety-Constrained Training	Performance-Focused	Safety Prioritization	Medium
Verification-Guided Training	Post-Development Verification	Integrated Verification	High
Robust Training	Nominal Conditions	Worst-Case Scenarios	Medium

Table 4: Bridging Regulatory Gaps for AI Certification [9, 10]

6. Explainable AI for Safety Validation

Explainability techniques have evolved from academic curiosities into essential safety assurance components for AI systems. Neural networks' "black box" nature presents fundamental safety validation challenges, as traditional verification approaches rely on understanding system behavior to identify potential failure modes. Explainable AI (XAI) techniques address this challenge, providing insights into neural networks' internal decision processes, enabling more rigorous safety validation, and supporting certification efforts in regulated domains [11].

Causal analysis methods identify inputs most strongly influencing safety-critical decisions, providing essential safety validation insights. These techniques quantify relationships between specific inputs and model outputs, highlighting features driving critical decisions. Research on explanation in artificial intelligence demonstrates that understanding causality remains fundamental to human explanations, and applying similar approaches to AI systems reveals critical dependencies requiring validation for safety-critical applications. By revealing these causal relationships, verification efforts focus on most safety-relevant system behavior aspects [11].

Safety-focused attention mechanisms highlight rationales behind AI decisions, making neural network behavior more transparent and amenable to safety analysis. These mechanisms visualize regions or features that networks focus on when making decisions, providing insights into underlying reasoning processes. Social science research on explanations shows that selective highlighting of relevant features constitutes key components of effective explanations, inspiring similar approaches for neural networks supporting safety validation by confirming appropriate decision factors [11].

Symbolic rule extraction from trained networks enables traditional safety analysis by translating complex neural network behavior into interpretable rules. These approaches distill knowledge embedded in neural networks into symbolic representations analyzable using conventional safety engineering techniques. Research drawing from social sciences emphasizes the importance of contrastive explanations, highlighting why one decision was made instead of another, informing rule extraction approaches capturing critical distinctions for safety analysis [11].

7. Conclusion

The fundamental transformation in the philosophy of functional safety in the AI-enabled embedded systems signifies the paradigm shift in safety engineering, no longer based on deterministic systems, but a multi-layer structure with a probabilistic nature of machine learning incorporated. With the integration of design-time verification, runtime monitoring, safety-aware training, as well as hybrid architecture, researchers have introduced a possible route to obtaining AI capabilities in safety-critical applications. These methods recognize that complex neural networks have inherent limitations in verification and put in countervailing measures which will preserve safety, through verification backup measures and constant monitoring. When these methodologies evolve and are accepted by the regulators, they will allow the use of AI in highly regulated areas at scale without necessarily a decline in safety outcomes. Combined with the fact that the above approaches are complementary, in that formal verification offers guarantees where feasible, runtime monitoring identifies likely violations at operation time, and hybrid safety architectures use safety boundaries, they yield safety frameworks that are resilient to the particular demands of AI-enabled systems. Such an evolution in

functional safety processing is only going to persist when AI systems become more autonomous, necessitating ever-more-advanced safety paradigms to allow safe and dependable operation within a wide range and complex operating conditions.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Dario Guidotti, "Verification of Neural Networks for Safety and

Security-critical Domains," University of Sassari, Piazza Università 21, 2022. [Online]. Available: https://ceur-ws.org/Vol-3345/paper10_RiCeRCa3.pdf

[2]Mohd Hafeez Osman et al., "Run-Time Safety Monitoring Framework for AI-Based Systems: Automated Driving Cases," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/338364338_Run-Time_Safety_Monitoring_Framework_for_AI-Based_Systems_Automated_Driving_Cases

[3] PeterMcCluskey, "Provably Safe AI," LessWrong, 2023. [Online]. Available: https://www.lesswrong.com/posts/KX3Qwr7QM7CvhJLG6/provably-safe-ai

[4] Hong Wang et al., "A Survey on an Emerging Safety Challenge for Autonomous Vehicles: Safety of the Intended Functionality," Engineering, Volume 33, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809924000274

[5] Hira Naveed et al., "Towards Runtime Monitoring for Responsible Machine Learning

using Model-driven Engineering," ACM, 2024. [Online]. Available: https://nzjohng.github.io/publications/papers/models2024.pdf

[6] Xiaoyu Zhang et al., "Deep Learning Library Testing: Definition, Methods and Challenges," ACM Computing Surveys, Volume 57, Issue 7, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3716497

[7] Leonardo de Moura and Nikolaj Bjørner, "Satisfiability Modulo Theories: Introduction and Applications," Communications of the ACM, 2011. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/1995376.1995394

[8] Guy Katz et al., "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," arXiv:1702.01135, 2017. [Online]. Available: https://arxiv.org/abs/1702.01135

[9] Rick Salay et al., "An Analysis of ISO 26262: Using Machine

Learning Safely in Automotive Software," arXiv:1709.02435v1, 2017. [Online]. Available: https://esnl.hnu.edu.cn/ecps2018/Using_Machine_Learning_Safely_in_Automotive_Software.pdf

[10] Simon Burton et al., "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," Artificial Intelligence, Volume 279, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370219301109

[11] Tim Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, Volume 267, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988

[12] Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry, "Formal methods for semi-autonomous driving," 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), 2015. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7167334