IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Science, Engineering, and Applied Research

Vol 1, No 1, 2026

Trustworthy Artificial Intelligence in Safety-Critical Engineering Systems: A Theoretical Framework of Verification, Robustness and Certification for Learned Components

Atolagbe Umar Kawu[1], NWOGBOJI Emmanuel Usulor[2], OGAH SUNDAY OGBO[3], Igwe Emmanuel Sunday[4], Paul Idowu Bamigbola[5]

[1,2,3,4,5] Veristzon Global Research Lab

ABSTRACT

Machine-learning components are increasingly incorporated into systems whose failure can cause death: vehicles, aircraft, vessels, medical devices, industrial control and power systems. Established safety engineering assures such systems through three steps, specifying what the system must do, verifying the system against that specification, and obtaining certification from an authority, and learned components resist all three. Their behaviour is learned rather than specified, tested rather than verified, and liable to change when they are updated or when the conditions they encounter shift. This paper develops a framework for assuring safety-critical systems that contain learned components, drawing on safety and reliability engineering, the formal-verification and assurance-case literatures, and recent research on the robustness, verification and certification of machine-learning systems. It distinguishes three assurance problems: the specification problem, that what the component should do cannot be fully stated; the verification problem, that the component cannot be fully shown to do it; and the stability problem, that the component cannot be assumed to continue doing it after an update or a change in operating conditions. Five responses are evaluated, constrained architectures, runtime monitoring with fallback, statistical verification, formal methods for bounded properties, and assurance cases that argue rather than prove, and matched to classes of system and failure. The paper concludes by considering the implications for engineering education, professional responsibility and standards bodies in jurisdictions such as Nigeria, which will import most such systems rather than certify them. It argues that assurance for learned components must shift from proving that a component meets a fixed specification to arguing, on monitored evidence, that it remains within acceptable bounds, and that jurisdictions which import rather than certify these systems must build the capacity to evaluate the assurance cases that accompany them rather than relying on certifications issued elsewhere.

Keywords: safety-critical systems; machine learning; verification and validation; robustness; certification

Trustworthy Artificial Intelligence in Safety-Critical Engineering Systems: A Theoretical Framework of Verification, Robustness and Certification for Learned Components

1. INTRODUCTION

Safety engineering rests on a straightforward principle: a system whose failure could cause death or serious harm should not be deployed until there is adequate confidence that it will behave safely. The discipline has developed elaborate means of establishing that confidence. Hazards are identified and analysed, safety requirements are specified, designs are verified against those requirements by analysis, inspection and test, and independent authorities certify that the evidence is sufficient. Refined over decades in aviation, rail, nuclear power, automotive engineering and medical devices, these practices have made many safety-critical technologies remarkably safe.

Machine learning challenges these practices at their foundations. A neural network that detects pedestrians, a reinforcement-learning controller that steers a vessel, or a model that interprets sensor data in an industrial plant is not designed to a specification in the conventional sense. Its behaviour is learned from data; its internal logic cannot be read or reasoned about as conventional software can; its performance is demonstrated by testing on samples rather than proved for all inputs; and its behaviour may change when it is retrained or when the conditions it encounters drift from those on which it was trained. Demonstrating the safety of machine learning has accordingly been described as one of the most challenging obstacles to its deployment in safety-critical applications (Burton & Herd, 2023; Goodloe, 2023).

This paper develops a theoretical framework for assuring safety-critical systems that contain learned components. It distinguishes three assurance problems, specification, verification and stability, evaluates five classes of response, and considers which combinations of response are appropriate to which systems and failures. It then examines what the framework implies for engineering practice in countries such as Nigeria, which will for the most part import rather than develop and certify AI-enabled safety-critical systems.

The analysis is theoretical, integrating recent research on the safety assurance of machine learning (Burton & Herd, 2023; Dong et al., 2023; Goodloe, 2023; McCutcheon et al., 2026; Paterson et al., 2025), on hazard analysis and the validation of autonomous systems (Feng et al., 2023; Molloy et al., 2024), on runtime assurance and safety filters (Peng et al., 2023; Sheikhi et al., 2025; Vaaler et al., 2024), on formal and statistical verification (Lindemann et al., 2025; Newton & Papachristodoulou, 2023; Rossi et al., 2024), on certification (Neumaier et al., 2025), on adaptation after deployment (Abad et al., 2026; Ballingall et al., 2023), and on AI governance and professional responsibility (Doyle et al., 2026; Nguyen, 2026; Okon, 2026).

2. LITERATURE REVIEW

2.1 The Established Assurance Paradigm

Conventional safety assurance proceeds through a chain of activities. Hazard analysis identifies the ways in which a system could cause harm; safety requirements state what the system must and must not do to avoid those harms; design and implementation are traced to the requirements; verification demonstrates, through analysis, review and testing, that the implementation satisfies the requirements; and certification by a regulator or independent assessor judges whether the evidence is adequate for the level of risk. For software, this chain is codified in domain standards that demand more rigorous development and verification for functions whose failure would be more severe.

The paradigm rests on three assumptions. The first is that the intended behaviour can be specified, that it is possible to state what the system should do in each relevant situation. The second is that the implementation can be verified against that specification, whether by exhaustive analysis or by testing which, combined with structural coverage and design review, affords confidence that the implementation behaves as specified across its input space. The third is that the verified behaviour is stable: once certified, the system continues to behave as it did, and any change triggers re-verification.

The paradigm is increasingly organised around assurance cases, structured arguments, supported by evidence, that a system is acceptably safe for a given application in a given context, which make explicit the claims advanced, the evidence offered for them and the reasoning connecting the two. The emerging consensus in industry and academia holds that assurance cases are also the appropriate vehicle for assuring learning-enabled systems (Dong et al., 2023; Paterson et al., 2025). The difficulty is that

© IIARJournal. All rights reserved.

each of the paradigm's three assumptions fails, at least in part, for learned components, which makes the arguments in such cases harder to construct and the supporting evidence harder to obtain.

3. CONCEPTUAL FRAMEWORK

3.1 Three Assurance Problems

The failure of the three assumptions gives rise to three corresponding problems. The first is a problem of specification. Machine learning is adopted precisely where the intended behaviour is difficult to specify: no one can write a complete rule for what constitutes a pedestrian in every condition of lighting, pose and occlusion, which is why perception systems learn the concept from labelled examples. The training data and the learning objective serve in place of a specification, but they constitute an incomplete and implicit one, the data cover some situations and not others, the labels reflect choices and errors, and the objective rewards average performance rather than correct behaviour in each critical case. Hazard analysis of machine-learning-based perception in autonomous vehicles must therefore contend with the fact that the perception function's intended behaviour is defined only partly and implicitly (Molloy et al., 2024). This has a direct consequence for assurance: if the intended behaviour cannot be fully stated, it cannot be fully verified, since verification requires a standard against which to verify. Assurance must consequently work with partial specifications, expressed as properties that can be stated, that the output remains within physical bounds, that small perturbations of the input do not change the output drastically, or that the system detects when an input lies outside its competence, supplemented by evidence of performance on representative data. Burton and Herd (2023) identify uncertainty in the specification of the operating context and of the required behaviour as a central source of difficulty in assuring machine learning.

The second is a problem of verification. Even where properties can be specified, demonstrating that a learned component satisfies them is difficult. Testing on samples establishes how the component behaves on those samples but not how it behaves elsewhere, and the input spaces of perception and control functions are vast. Formal verification can, in principle, establish properties for all inputs within a region, through methods such as interval arithmetic combined with formal provers for closed-loop systems (Rossi et al., 2024) and polynomial optimisation for verifying neural networks (Newton & Papachristodoulou, 2023); but these methods scale poorly to large networks and can verify only bounded, precisely stated properties. The validation of autonomous systems faces a further obstacle in that the events relevant to safety are rare, so that a vehicle might require testing over enormous distances to encounter enough safety-critical situations to estimate its failure rate with confidence. Feng et al. (2023) address this through dense reinforcement learning, which trains an adversarial environment to generate safety-critical scenarios more frequently, substantially accelerating validation while preserving the ability to estimate real-world risk; such methods render statistical verification more feasible, but they do not convert it into proof. Reliability assessment offers a means of making statistical evidence usable in assurance: Dong et al. (2023) develop methods for assessing the reliability of machine-learning components and for constructing safety arguments from that assessment, connecting operational profiles, robustness evidence and reliability claims within an assurance case and accepting that verification will be probabilistic while seeking to make the probabilistic claims rigorous.

The third is a problem of stability, the possibility that a learned component's behaviour will not remain as it was assured. Two mechanisms produce such instability. The first is distribution shift: the conditions the component encounters in operation may differ from those represented in its training and test data, and performance may degrade with no change to the component itself. Research on machine learning for materials shows that apparent tests of generalisation frequently probe interpolation rather than genuine extrapolation, so that performance on genuinely novel inputs may be considerably worse than standard evaluations suggest (Li et al., 2025). The second mechanism is update: learned components may be retrained on new data to improve performance or correct errors, but each update may alter behaviour in ways that invalidate previous assurance. The stability problem is especially acute for systems designed to adapt. Ballingall et al. (2023), interviewing experts on automated driving systems that can adapt using machine learning, document the difficulty of assuring safety when behaviour may change after deployment, and research on adapting deployed AI systems under operational constraints reviews methods for modifying models while preserving properties of the original (Abad et al., 2026). The

© IIARJournal. All rights reserved.

central question nevertheless persists: how much evidence is required to accept an update, and how can that evidence be produced quickly enough for the update to be of use?

3.2 Five Responses to the Assurance Problems

The literature offers five broad responses to these problems, each addressing some more effectively than others. The first is to constrain the role the learned component plays within the system, so that its failures cannot directly cause harm. The learned component may be confined to non-safety functions, paired with a conventional component that checks its outputs, or bounded by limits enforced through verified logic. Neumaier et al. (2025) propose validating a machine-learning model for certification by means of symbolic regression and a behaviour envelope, so that the model's behaviour can be characterised and bounded in terms an assessor can examine. Constrained architectures address the specification and verification problems by reducing what must be specified and verified about the learned component itself, at the cost of possibly limiting the performance the component can contribute.

The second response accepts that the learned component cannot be fully verified and instead monitors it during operation, switching to a verified fallback when the monitor detects unsafe conditions. The classical form is the simplex architecture, in which a high-performance but unverified controller operates normally and a verified safety controller assumes control as the system approaches unsafe states. Sheikhi et al. (2025) extend this to multi-agent cyber-physical systems through a black-box simplex architecture; Peng et al. (2023) propose a runtime-assurance framework for behaviour planning that combines imitation learning with a formal safety model; and Vaaler et al. (2024) develop a modular control architecture for marine navigation in which predictive safety filters constrain a reinforcement-learning controller. Monitoring may also be directed at the inputs, with out-of-distribution detection and certainty scoring identifying inputs on which the learned component is likely to be unreliable (Berenbeim et al., 2025). Runtime monitoring addresses all three problems in part, but its own correctness must be assured, and its effectiveness depends upon the availability of a safe fallback.

The third response gathers statistical evidence of performance and reliability and reasons from it to a quantified safety claim. Accelerated testing (Feng et al., 2023), reliability assessment (Dong et al., 2023) and conformal prediction, which provides distribution-free statistical guarantees on prediction errors, belong to this class; Lindemann et al. (2025) show how conformal prediction can be combined with formal verification and control to yield practical safety guarantees for autonomous systems. Statistical verification addresses the verification problem directly, but its guarantees hold only under assumptions about the relationship between test and operating conditions, which renders it vulnerable to the stability problem. The fourth response applies formal verification to specific, precisely stated properties of learned components or of the closed-loop systems containing them (Newton & Papachristodoulou, 2023; Rossi et al., 2024). Formal methods furnish the strongest guarantees for the properties they address, but they address only properties that can be stated precisely and they scale poorly; their most effective use is therefore to verify critical bounded properties, that a controller never commands an actuator beyond its limits, or that a classifier's output is stable within a neighbourhood of an input, rather than to verify complete behaviour.

The fifth response integrates the others within a structured argument. An assurance case for a learning-enabled system makes explicit what is claimed about the system's safety, what evidence supports each claim, and what assumptions the argument depends upon. Paterson et al. (2025) develop a methodology for the safety assurance of machine learning in autonomous systems that addresses the stages of the machine-learning lifecycle, from requirements and data management through model learning and verification to deployment; Hodge and Osborne (2025) adapt such a methodology to agile development so that assurance can keep pace with iterative development; and McCutcheon et al. (2026) argue for the integration of human, software and machine-learning assurance paradigms, since AI-enabled systems combine characteristics that each paradigm alone handles poorly. Assurance cases do not resolve the three problems; rather, they make visible how each has been addressed and where residual uncertainty remains, which is precisely what a certifying authority requires in order to judge whether the risk is acceptable.

3.3 Matching Responses to Systems and Failures

© IIARJournal. All rights reserved.

The five responses are not interchangeable, and the appropriate combination depends on the role the learned component plays and on the character of its potential failures. Three roles may be distinguished. In an advisory role, a learned component informs a human or conventional decision-maker who remains responsible for action, as when a model flags anomalies in plant data for an operator. In a perception role, it interprets the environment for a downstream controller, as when a neural network detects obstacles for an automated vehicle (Molloy et al., 2024). In a control role, it directly commands actuators, as when a reinforcement-learning policy steers a vessel (Vaaler et al., 2024).

Failures differ along two dimensions that bear on assurance. The first is severity, whether a failure causes inconvenience, damage or death. The second is recoverability, whether there is time and opportunity to detect a failure and reach a safe state before harm occurs. A failure in an advisory role is usually recoverable, since a human stands between the output and the action, although automation bias can erode that protection; a failure in perception may be recoverable if downstream logic can detect implausible outputs and unrecoverable if it cannot; and a failure in direct control may leave no time for recovery at all.

These distinctions suggest a pattern. Advisory components with recoverable failures may be assured chiefly through statistical evidence of performance together with monitoring of their use (Dong et al., 2023). Perception components whose failures feed safety-critical decisions require statistical verification supplemented by out-of-distribution detection and by downstream plausibility checks capable of rejecting anomalous outputs (Berenbeim et al., 2025; Feng et al., 2023). Control components with severe and unrecoverable failure modes require the strongest combination: architectural constraints and runtime safety filters that keep commanded actions within verified bounds, together with formal verification of the bounding properties (Lindemann et al., 2025; Rossi et al., 2024; Sheikhi et al., 2025). In every case the assurance case should state which role the component plays, which failures it can cause, and which responses address them (Paterson et al., 2025).

3.4 Principal Propositions of the Framework

The argument may be summarised in a set of propositions open to research and usable as design principles for engineers, assessors and regulators.

The first is foundational: the assurance of learned components faces three distinct problems, specification, verification and stability, so that an assurance strategy addressing only verification, for instance through extensive testing, will leave the specification and stability problems unaddressed (Burton & Herd, 2023; Goodloe, 2023). From this follows a proposition concerning architecture: the more severe the consequences of a learned component's failure, the more its role should be constrained by architecture, so that its failures are bounded by verified components; learned components should not constitute the sole barrier against catastrophic hazards (Neumaier et al., 2025; Sheikhi et al., 2025).

A further group of propositions concerns the responses themselves. Runtime monitoring with a verified fallback is the most general response to the stability problem, because it addresses distribution shift and update at the point of operation, though its value depends on the reliability of the monitor and the availability of a safe fallback state (Berenbeim et al., 2025; Peng et al., 2023; Vaaler et al., 2024). Statistical verification provides quantified evidence of safety only under assumptions about the similarity of test and operating conditions, so that its guarantees must be accompanied by monitoring of those assumptions in operation (Dong et al., 2023; Feng et al., 2023; Lindemann et al., 2025). Formal methods should be directed at critical bounded properties rather than at complete behaviour, where their guarantees are strongest and their limitations in scalability least binding (Newton & Papachristodoulou, 2023; Rossi et al., 2024).

A proposition concerning adaptation holds that systems which adapt after deployment require assurance processes capable of evaluating updates as rigorously as initial certification, and at the pace of the updates; where this cannot be achieved, adaptation in safety-critical functions should be restricted (Abad et al., 2026; Ballingall et al., 2023; Hodge & Osborne, 2025). Finally, assurance cases are the appropriate vehicle for integrating the responses, since they make explicit which problems have been addressed, by what evidence and with what residual uncertainty; their adequacy depends upon the competence and independence of those who construct and assess them (McCutcheon et al., 2026; Paterson et al., 2025).

© IIARJournal. All rights reserved.

4. ANALYSIS AND DISCUSSION

4.1 Importing Jurisdictions and the Nigerian Case

Most of the assurance literature presupposes a jurisdiction that develops and certifies safety-critical systems, with domestic manufacturers, specialised regulators and a technical community able to construct and assess assurance cases. Many countries, including Nigeria, occupy a different position, importing most AI-enabled safety-critical systems, vehicles with driver-assistance and automated functions, aircraft and avionics, medical devices with AI components, industrial control systems for oil, gas and power, and drones, after they have been assured and certified, if at all, by authorities elsewhere.

This position gives rise to three distinctive challenges. The first concerns the transfer of assurance. An assurance case is valid for a given application in a given context, and a system certified for European roads, North American hospitals or temperate climates has been assured under assumptions about its operating environment that may not hold in Nigeria, where road users behave differently, infrastructure differs, climate, dust and power quality differ, and maintenance practices differ. The stability problem thus becomes a problem of transfer, since a component whose behaviour was assured under one distribution of conditions may behave differently under another (Li et al., 2025); imported assurance cannot be accepted uncritically, and the assumptions on which it rests must be examined against local conditions.

The second challenge concerns regulatory capacity. Assessing whether imported assurance is adequate for local conditions requires regulators with technical competence in machine learning and safety assurance, a competence scarce everywhere and especially so in regulators with limited resources. Since AI is altering both what regulators govern and the informational environment within which they govern, regulatory capacity must adapt (Doyle et al., 2026), and research on AI governance and public policy in Nigeria documents challenges of institutional capacity, infrastructure and policy coherence (Okon, 2026). For safety-critical engineering systems, regulators need not reproduce full certification, but they do require the capacity to examine the operating-context assumptions of imported assurance, to require evidence that those assumptions hold locally, and to mandate runtime monitoring and the reporting of incidents in local operation.

The third challenge concerns professional responsibility. Engineers in importing jurisdictions will specify, procure, integrate, operate and maintain AI-enabled systems they did not design, and their professional responsibility does not lapse because the learned components were assured elsewhere. An engineer who integrates an imported perception system into a local application, or who operates an AI-enabled control system in a Nigerian plant, is responsible for judging whether the system is safe in that context, a judgement that requires understanding the three assurance problems well enough to put the right questions to suppliers and to recognise when local conditions fall outside the assured envelope. As professional ethics education moves from a focus on legal compliance toward a broader social responsibility in the age of AI (Nguyen, 2026), that responsibility, for engineering, extends to judging the safety of learned components in context.

A practical strategy for importing jurisdictions follows. Procurement and approval should require suppliers to disclose the operating-context assumptions of their assurance cases and to provide evidence relevant to local conditions; deployment should favour systems with constrained architectures and runtime monitoring, so that the effects of distribution shift are detected and bounded in local operation (Berenbeim et al., 2025; Sheikhi et al., 2025); incident reporting and post-deployment monitoring should be mandated, so that evidence about local performance accumulates; and engineering education and professional development should equip engineers to engage critically with assurance cases rather than to accept them as given.

4.2 Implications

For engineers designing safety-critical systems with learned components, the framework implies that assurance should be planned from the outset around all three problems. Specification should identify the properties that can be stated and the operating context within which they must hold; verification should combine statistical evidence, formal verification of bounded properties and architectural constraints; and stability should be addressed through runtime monitoring and controlled update processes, with the assurance case making explicit how each problem has been addressed (Paterson et

© IIARJournal. All rights reserved.

al., 2025). For assessors and regulators, the implication is that the evaluation of learning-enabled systems should examine the specification and stability problems as closely as the verification problem: an assurance case built mainly on test results should be scrutinised for the representativeness of its test data, the explicitness of its operating-context assumptions, and the provisions for detecting and responding to distribution shift (Burton & Herd, 2023; Dong et al., 2023), and regulators should develop processes for evaluating updates to adaptive systems (Ballingall et al., 2023).

For engineering education, the framework supports curricula that teach the safety assurance of learned components alongside machine learning itself, so that engineers who build or integrate AI-enabled systems understand why testing is not proof, why distribution shift matters, how runtime monitoring operates, and how assurance cases are constructed and challenged (Goodloe, 2023; Nguyen, 2026); in Nigeria and comparable jurisdictions, such education is a precondition for the critical engagement with imported systems that the framework recommends. For standards bodies and professional institutions, including those in jurisdictions that mainly import, there is a corresponding role in developing guidance on the local acceptance of imported AI-enabled systems, covering the evidence suppliers should provide, the local validation required, and the monitoring and reporting obligations of operators (Doyle et al., 2026; Okon, 2026).

Several questions remain open for research: how operating-context assumptions can be stated precisely enough to be checked in a new jurisdiction; how effectively runtime monitors detect the kinds of distribution shift that arise when systems are transferred across countries; what evidence suffices to accept an update to an adaptive safety-critical system; and how importing jurisdictions can build assurance capacity efficiently. Answering these questions requires collaboration between machine-learning researchers, safety engineers and regulators, and evidence drawn from the jurisdictions in which imported systems are actually used.

5. CONCLUSION AND RECOMMENDATION

Learned components are entering systems whose failure can cause death, and the established practices of safety engineering were not designed for them. This paper has argued that their assurance faces three distinct problems: their intended behaviour cannot be fully specified, their conformity cannot be fully verified, and their behaviour cannot be assumed to remain stable after update or under changed conditions. No single response resolves all three. Constrained architectures limit what must be assured; runtime monitoring with a fallback detects and bounds failures in operation; statistical verification quantifies performance under stated assumptions; formal methods prove bounded properties; and assurance cases integrate these responses and make residual uncertainty visible.

For jurisdictions such as Nigeria, which will for the most part import AI-enabled safety-critical systems, the framework identifies a particular risk: assurance produced elsewhere rests on assumptions about operating conditions that may not hold locally, and the stability problem becomes a problem of transfer. Meeting this risk does not require importing jurisdictions to replicate full certification; it requires the capacity to examine the assumptions of imported assurance, to favour architectures that detect and bound failures in local operation, to mandate monitoring and incident reporting, and to educate engineers capable of engaging critically with the assurance of the systems they integrate and operate.

Trustworthy AI in safety-critical engineering will not arise from more powerful learning methods alone. It will arise from assurance practices that are candid about what can and cannot be demonstrated, from architectures that keep learned components within verified bounds, and from engineers and regulators, in every jurisdiction where such systems operate, who understand the limits of the evidence on which their safety depends.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

The paper contributes a decomposition of the assurance of artificial intelligence in safety-critical systems into three problems that the established assurance paradigm does not resolve, together with five responses and a mapping of responses to systems and failure types.

It gives engineers and regulators, particularly in jurisdictions that import safety-critical AI rather than develop it, a basis for assuring systems whose behaviour cannot be fully specified in advance, and for matching the assurance response to the kind of system and the kind of failure at stake.

© IIARJournal. All rights reserved.

7. REFERENCES

Abad, M., Nin, J., & Unceta, I. (2026). Adapting deployed AI systems under operational constraints: A review of differential replication methods. Machine Learning with Applications, 25, 100948. https://doi.org/10.1016/j.mlwa.2026.100948

Ballingall, S., Sarvi, M., & Sweatman, P. (2023). Safety assurance for automated driving systems that can adapt using machine learning: A qualitative interview study. Journal of Safety Research, 84, 243–250. https://doi.org/10.1016/j.jsr.2022.10.024

Berenbeim, A. M., Cobb, A. D., Roy, A., Jha, S., & Bastian, N. D. (2025). Applications of certainty scoring for machine learning classification and out-of-distribution detection. ACM Transactions on Probabilistic Machine Learning, 1(4), 1–35. https://doi.org/10.1145/3711713

Burton, S., & Herd, B. (2023). Addressing uncertainty in the safety assurance of machine-learning. Frontiers in Computer Science, 5, 1132580. https://doi.org/10.3389/fcomp.2023.1132580

Dong, Y., Huang, W., Bharti, V., Cox, V., Banks, A., Wang, S., Zhao, X., Schewe, S., & Huang, X. (2023). Reliability assessment and safety arguments for machine learning components in system assurance. ACM Transactions on Embedded Computing Systems, 22(3), 1–48. https://doi.org/10.1145/3570918

Doyle, C., Cave, M., & Webb, W. (2026). Artificial intelligence, regulatory capacity, and the adaptive state. Regulation & Governance. Advance online publication. https://doi.org/10.1111/rego.70210

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., & Liu, H. X. (2023). Dense reinforcement learning for safety validation of autonomous vehicles. Nature, 615(7953), 620–627. https://doi.org/10.1038/s41586-023-05732-2

Goodloe, A. E. (2023). Assuring safety-critical machine learning-enabled systems: Challenges and promise. Computer, 56(9), 83–88. https://doi.org/10.1109/mc.2023.3266860

Hodge, V. J., & Osborne, M. (2025). Agile development for safety assurance of machine learning in autonomous systems (AgileAMLAS). Array, 27, 100482. https://doi.org/10.1016/j.array.2025.100482

Li, K., Rubungo, A. N., Lei, X., Persaud, D., Choudhary, K., DeCost, B., Dieng, A. B., & Hattrick-Simpers, J. (2025). Probing out-of-distribution generalization in machine learning for materials. Communications Materials, 6(1), 9. https://doi.org/10.1038/s43246-024-00731-w

Lindemann, L., Zhao, Y., Yu, X., Pappas, G. J., & Deshmukh, J. V. (2025). Formal verification and control with conformal prediction: Practical safety guarantees for autonomous systems. IEEE Control Systems, 45(6), 72–122. https://doi.org/10.1109/mcs.2025.3611545

McCutcheon, R., Joiner, K. F., Qiao, L., & Harvey, J. (2026). Assurance of safety-critical AI-enabled capabilities: Integrating human, software, and machine learning assurance paradigms. AI, 7(10), 435. https://doi.org/10.3390/ai7100435

Molloy, J., Shahbeigi, S., & McDermid, J. A. (2024). Hazard and safety analysis of machine-learning-based perception capabilities in autonomous vehicles. Computer, 57(11), 60–70. https://doi.org/10.1109/mc.2024.3443751

Neumaier, M., Anselment, M., & Rudolph, S. (2025). Validation of a machine learning model for certification using symbolic regression and a behaviour envelope. Aerospace, 12(5), 412. https://doi.org/10.3390/aerospace12050412

Newton, M., & Papachristodoulou, A. (2023). Sparse polynomial optimisation for neural network verification. Automatica, 157, 111233. https://doi.org/10.1016/j.automatica.2023.111233

Nguyen, Q. K. (2026). Professional ethics education in the age of artificial intelligence: From legal compliance to social responsibility. International Journal of Ethics Education. Advance online publication. https://doi.org/10.1007/s40889-026-00257-4

Okon, U. (2026). Artificial intelligence, governance, and public policy in Nigeria: Challenges, opportunities and policy imperatives. International Journal of Public Administration, 5(2), 97–110. https://doi.org/10.4314/ijopad.v5i2.6

Paterson, C., Hawkins, R., Picardi, C., Jia, Y., Calinescu, R., & Habli, I. (2025). Safety assurance of machine learning for autonomous systems. Reliability Engineering & System Safety, 264, 111311. https://doi.org/10.1016/j.ress.2025.111311

Peng, Y., Tan, G., & Si, H. (2023). RTA-IR: A runtime assurance framework for behavior planning based on imitation learning and responsibility-sensitive safety model. Expert Systems with Applications, 232, 120824. https://doi.org/10.1016/j.eswa.2023.120824

Rossi, F., Bernardeschi, C., & Cococcioni, M. (2024). Neural networks in closed-loop systems: Verification using interval arithmetic and formal prover. Engineering Applications of Artificial Intelligence, 137, 109238. https://doi.org/10.1016/j.engappai.2024.109238

Sheikhi, S., Mehmood, U., Bak, S., Smolka, S. A., & Stoller, S. D. (2025). The black-box simplex architecture for runtime assurance of multi-agent CPS. Innovations in Systems and Software Engineering, 21(2), 619–634. https://doi.org/10.1007/s11334-024-00553-6

© IIARJournal. All rights reserved.

Vaaler, A., Husa, S. J., Menges, D., Larsen, T. N., & Rasheed, A. (2024). Modular control architecture for safe marine navigation: Reinforcement learning with predictive safety filters. Artificial Intelligence, 336, 104201. https://doi.org/10.1016/j.artint.2024.104201

© IIARJournal. All rights reserved.