1

Edge AI Systems Verification and Validation

Ovidiu Vermesan[1], Alain Pagani[2], Roy Bahr[1], Marcello Antonio Coppola[3], and Giulio Urlini[4]

[1]SINTEF AS, Norway
[2]German Research Center for Artificial Intelligence (DFKI), Germany
[3]STMicroelectronics, France
[4]STMicroelectronics, Italy

Abstract

The integration of edge artificial intelligence (AI) into different complex systems presents unique challenges, particularly concerning their reliability, robustness, safety, and transparency. Edge AI systems must function as intended and meet regulatory and technical standards. Traditional verification and validation (V&V) methodologies, which are well-suited for conventional software (SW) and hardware (HW) systems, do not fully address the unique characteristics of edge AI-based systems that include hardware, software, elements of edge AI technology stack and data.

The chapter delves into the challenges and methodologies for edge AI verification and validation to identify the unique elements required to develop verifiable edge AI systems based on a structured verification and validation framework integrated with model- and data-driven engineering principles, assurance cases, and domain-specific requirements. It highlights the terminology and concepts for edge AI as a technology that integrates HW, SW, and edge AI technology and data while presenting the challenges of the convergence of these technologies in developing verification and validation solutions.

Keywords: edge AI, edge AI system, verification, validation, machine learning, deep learning, AI agents, agentic AI, system engineering, small language models.

1.1 Introduction and Background

Edge AI has become a cornerstone of innovation in various industries, driving advancements in automation, decision-making, and predictive analysis. Edge AI systems applying machine learning (ML), deep learning (DL), and data processing at the edge involving deep neural networks (DNN) present significant challenges for ensuring the reliability, safety, and effectiveness of intelligent embedded devices across the edge AI computing continuum, ranging from micro- to deep- and meta-edge. Edge AI can be either deterministic or non-deterministic, based on the typical application and design choices involved. Many edge AI applications prioritise real-time, deterministic behaviour for critical tasks, such as control algorithms. Other applications can leverage the non-deterministic nature of AI to deliver more adaptable and creative solutions as the non-deterministic nature of edge AI means it can offer different interpretations based on context. In real-time applications, edge AI systems require precise timing and consistent response times. This is demanded for tasks where milliseconds of delay can be critical. Deterministic edge AI is appropriate for applications that demand predictability and consistency, while non-deterministic approaches are advantageous for applications that require adaptability, creativity, and continuous learning. The choice between using a deterministic or non-deterministic approach finally depends on the detailed requirements of the application and the expected trade-offs among predictability, adaptability, and computational cost.

The advancement of edge AI technologies and the ubiquity of automated AI-based tools have created complex operational environments. Edge AI systems are evolving towards engineering advanced adaptive systems and require new concepts for verification and validation to address the challenging multidimensional integration of HW, SW, AI models, algorithms, datasets, and the multimodality of data.

The advantages of leveraging edge AI in many industrial applications include real-time processing, enhanced privacy and data security, reduced latency, optimised bandwidth, reliability, and scalability, as illustrated in Figure 1.1.

Edge AI technology stack combines AI and IoT with edge computing, allowing data processing and edge AI algorithm execution to occur directly on devices located at the edge of the network. By bringing AI closer to the source of data generation, edge AI enables more efficient and responsive decision-making across a wide range of applications.

AI systems, particularly those based on machine learning (ML), pose unique challenges that differ from traditional software. Unlike conventional

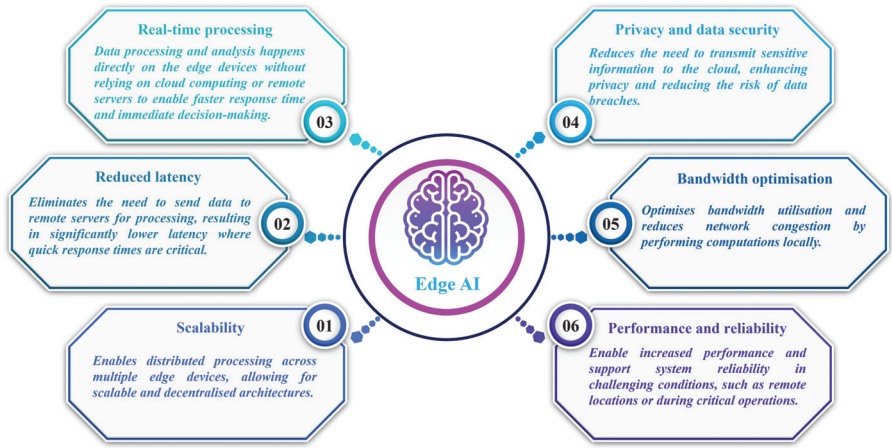

Figure 1.1 Edge AI advantages.

programs where behaviour is largely determined by explicit code, AI system behaviour often emerges from complex interactions between algorithms, vast datasets, and the operational environment [1].

Many advanced AI models, particularly deep neural networks, function as "black boxes," making their internal decision-making processes difficult to understand, predict, or inspect directly, hence difficult to validate [2]. This opacity, combined with potential determinism/non-determinism and sensitivity to data variations, complicates efforts to guarantee reliability, safety, and fairness [3].

A particularly demanding application domain of edge AI is real-time machine vision, which is critical in domains such as industrial robotics, autonomous navigation, and quality inspection. In these systems, the correctness and timeliness of visual perception directly influence physical actions, safety, and mission success. Their dependence on high-throughput, often noisy and non-reproducible visual data, and the need for ultra-low latency, makes their verification and validation particularly challenging under edge constraints.

The data-driven approach is based on systematically and algorithmically producing the best dataset to feed a given AI-based model, focusing on improving data quality and data governance to enhance the performance of a specific problem statement. Data-driven AI aims to improve data quality and outcomes by treating code as an unchangeable entity and dealing with labelling, augmenting, managing, and curating data. This is part of the

data preprocessing, emphasising an iterative AI lifecycle consisting of data collection, model training, and error analysis.

The model-driven approach is based on producing the best model for a given dataset and aims to build new models and algorithmic improvements to enhance performance. The model-driven edge AI focuses on improving code reflecting the edge AI model or algorithm to achieve adequate results from fixed datasets. Edge AI developers view the training datasets from which the code, model, or algorithm is learning as a collection of reference labels. The edge AI model is made to fit that labelled training data and assumes the training data is external to the edge AI development process.

In model-driven edge AI, the focus is on optimising an edge AI model, whereas in data-driven edge AI, the focus is on data quality improvement. In model-driven edge AI, the aim is to find the most suitable edge AI model or an optimisation technique for a given problem, whereas, in data-driven edge AI, the aim is to find inconsistencies in the collected data for a given problem. The two approaches require specific verification and validation solutions.

Validation, in the context of edge AI systems, moves beyond the verification by checking if a system was built according to its technical specifications to seek confirmation that the edge AI system is fit for its intended purpose and effectively meet the actual needs and expectations of its users and stakeholders within its specific operational environment.

This necessitates the implementation of rigorous verification and validation processes, underscoring the responsibility and accountability in the development and implementation of edge AI systems.

The growing complexity and societal impact of AI, edge AI and generative AI demand a shift from purely technical verification towards a more holistic validation approach. This approach must encompass not only functional correctness but also usability, ethical alignment, fairness, robustness in real-world conditions, and overall effectiveness in achieving desired outcomes [6].

Before the adoption of AI agents and agentic AI with the use of large language models (LLMs), the development of autonomous and intelligent agents was deeply rooted in foundational paradigms of AI, such as multi-agent systems and expert systems, which emphasise social action and distributed intelligence [13][28].

Small language models (SLMs) are designed to offer capabilities similar to LLMs but scaled to edge computing capabilities, such as reduced size, processing requirements, and memory size. SLMs contain fewer parameters (e.g., hundreds of millions to one billion) while still providing strong performance for specific tasks.

Agentic AI is a class of systems that extends the capabilities of traditional AI agents by enabling multiple intelligent entities to collaborate on pursuing goals through shared memory [18][20], structured communication [24][22][26], and dynamic role assignment [21].

Ethical and legal aspects and the requirements on explainability and interpretability can lead to system development decisions that do not solely attempt to optimize functional requirements such as accuracy and robustness. In this case, system design choices rely on trade-offs that should ideally be made consciously by system developers.

Agentic AI systems pose challenges in explainability and verifiability due to their distributed, multi-agent architecture. While interpreting the behaviour of a single language model powered by the agent is already non-trivial, this complexity is multiplied when multiple agents interact asynchronously through loosely defined communication protocols. Each agent may possess its memory, task objective, and reasoning path, resulting in compounded opacity where tracing the causal chain of a final decision or failure becomes exceedingly difficult. The lack of shared, transparent logs or interpretable reasoning paths across agents makes it highly difficult, if not impossible, to determine why a particular sequence of actions occurred or which agent initiated a misstep. Compounding this opacity is the absence of formal verification tools tailored for agentic AI. In traditional software systems, model checking and formal proofs offer bounded guarantees, while there exists no widely adopted methodology to verify that a multi-agent system comprising multiple large language model agents collaborating on tasks will perform reliably across all input distributions or operational contexts.

Validation, therefore, serves as a cornerstone for building trustworthy edge AI, systems that stakeholders can confidently rely upon to operate safely, effectively, and responsibly [7]. It directly addresses the widening gap observed between accelerating edge AI capabilities and lagging safety protocols.

The AI verification standardisation efforts within the edge AI community underscores a fundamental challenge: establishing justified confidence, or trust, in edge AI systems whose behaviour often emerges unpredictably. This inherent uncertainty and the potential for significant negative impact necessitate rigorous V&V processes.

V&V encompasses activities designed to ensure that an edge AI system not only meets its specified requirements but also fulfils its intended purpose safely and reliably in its operational context. While drawing upon established V&V principles from software and systems engineering, edge AI verification

and validation requires tailored approaches and methodologies to address its specific complexities.

The emphasis on "trustworthiness" in standards and frameworks like ISO/IEC TR 24028 directly reflects this imperative to build demonstrable confidence in AI systems [4][5].

This chapter provides a comprehensive overview of the verification and validation of edge AI systems. It examines definitions grounded in international standards, outlines the core elements subject to verification and validation, details the typical process steps involved, analyses the significant research challenges, explores contextual variations, discusses current research trends, and summarises future directions needed to advance the field.

1.2 Foundational Concepts and Edge AI Verification and Validation Taxonomy

In edge AI systems, the failure of an AI component can lead to overall system failure, highlighting the need for AI V&V. Components with AI capabilities are treated as subsystems. V&V is carried out both on the AI subsystem itself and on its interfaces with other parts of the overall system, just as with any other subsystem. That is, the high-level definitions of V&V remain unchanged for systems containing one or more AI components.

AI V&V challenges require approaches and solutions that go beyond those for conventional or traditional systems (those without AI elements). In the context of edge AI systems, AI components and subsystems need to be integrated into the systems engineering framework. This involves identifying the characteristics of AI subsystems that create challenges in their V&V, highlighting these challenges, and providing potential solutions while determining open areas of research in the V&V of edge AI subsystems.

Conventional SW/HW systems are engineered via three main phases, namely, requirements, design and V&V. These phases are applied to each subsystem and to the system under design.

Before the expansion of AI, ML, DL, and generative AI, research on V&V of neural networks addressed the adaptation of existing standards (e.g., IEEE Std 1012-Software Verification and Validation) and processed the augmentation of these standards to enable V&V and new techniques and lessons learned to solve the V&V issues for systems integrating AI components.

In all the adaptation and augmentation attempts, one of the challenges is data validation, as the data upon which AI depends should go through a form of V&V process. Data quality attributes that are important for edge AI

systems include accuracy, currency and timeliness, correctness, consistency, usability, security and privacy, accessibility, accountability, scalability, lack of bias, and coverage and representativeness of the state space. Data validation steps can include file validation, transformation validation, import validation, domain validation, aggregation rule and business validation.

AI-based systems follow a distinct lifecycle compared to traditional systems. For edge AI systems learning lifecycle, V&V activities occur throughout the lifecycle, as illustrated in Figure 1.2. The requirements allocated to the edge AI subsystem encompass both hardware and software (HW/SW), as well as the AI models and data that flow up to the system from the edge AI subsystem.

Verification refers to the set of the activities that ensure that the edge AI system implements the specific function, and the system is built right according to requirements.

Edge AI system verification is the process of checking that the edge AI system achieves its goal without any bugs. It is the process to ensure whether the developed edge AI system is right or not. It verifies whether the developed product fulfils the requirements. Verification is static testing. Verification means answering the question: are we building the edge AI system, right?

Edge AI verification and validation require approaches and solutions at data, model and system level beyond those for cloud AI and conventional systems. Edge AI lifecycle workflows require to combine the SW/HW engineering methods with the data and system level analysis.

Data quality attributes like accuracy, timeliness, correctness, consistency, usability, security, privacy, accessibility, accountability, scalability, lack of

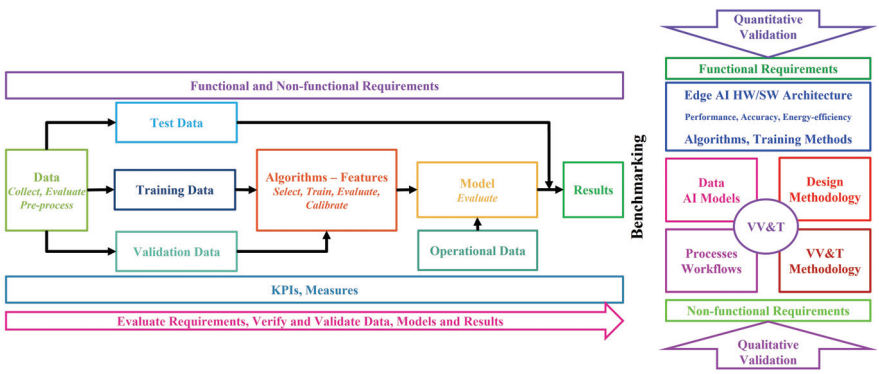

Figure 1.2 Edge AI verification and validation process.

bias, etc. are critical for edge AI. These data quality attributes are part of a larger edge AI non-functional requirements set.

Verification of edge AI systems involves systematically ensuring that AI models and their implementations fulfil specified requirements and intended purposes, as defined by recognised standards such as ISO/IEC 22989 (Information Technology - Artificial Intelligence - Concepts and Terminology). According to ISO, verification refers to the confirmation through objective evidence that specified requirements have been fulfilled. When applied to edge AI systems, verification processes ascertain that AI models and related software systems conform rigorously to technical and functional specifications, without necessarily validating the appropriateness of these specifications.

Principal elements involved in edge AI systems verification include:

- A formal requirements specification represents a crucial element, where clearly defined, unambiguous requirements serve as the foundational basis for verification. These specifications typically include functional requirements, performance criteria, safety constraints, security measures, and ethical guidelines.
- Model verification that entails evaluating AI models, including machine learning (ML) and deep neural networks (DNNs), ensuring their internal logic and behaviours align precisely with predefined specifications. Techniques employed in model verification include formal methods, theorem proving, model checking, and simulation-based testing.
- Software and hardware integration verification, which is vital, ensuring edge AI systems correctly interact with hardware components and software environments. It includes examining interface correctness, interoperability, real-time performance, and robustness under varying conditions and inputs.
- Rigorous test case generation and execution constitute essential verification steps. AI system verification employs automated test generation methods, including boundary value analysis, equivalence partitioning, and mutation testing, complemented by scenario-based testing to thoroughly assess compliance and performance under diverse and extreme operational conditions.
- Documentation and traceability processes involve detailed records demonstrating systematic compliance with verification steps, adherence to standards, and requirement fulfilment. Comprehensive documentation supports transparency and accountability and facilitates continuous improvement and iterative refinement processes.

In this context, the principal verification process involves several methodical steps:

- Requirement Analysis: Clearly define and document edge AI systems' functional, performance, and safety requirements.
- Verification Planning: Establishing a structured plan that details verification strategies, methods, criteria, and resources.
- Model and Code Inspection: Applying manual or automated inspections and formal verification techniques to analyse AI model structures and implementation code for correctness.
- Test Development: Generating extensive and varied test cases covering all possible usage scenarios, operational environments, and stress conditions.
- Verification Execution: Systematically conducting tests and verification activities, rigorously analysing outcomes against specified acceptance criteria.
- Reporting and Review: Documenting detailed verification outcomes, identifying discrepancies, and facilitating stakeholder review to ensure comprehensive verification coverage.
- Iterative Refinement: Addressing identified issues through iterative model adjustments, re-verification cycles, and continual improvement to achieve specified verification goals.

Validation refers to the set of the activities that ensure that the edge AI system that has been built is traceable to the requirements and the right edge AI system is built to meet user needs.

Validation is the process of checking whether the edge AI system is up to the mark or, in other words, if the product has high-level requirements. It is the process of checking the validation of the edge AI system, e.g., it checks if what we are developing is the right edge AI system. It is validation of the actual and expected edge AI systems. Validation is a form of dynamic testing. Validation means answering the question: are we building the right edge AI system?

Validation of edge AI systems is a critical and systematic process intended to ensure that the developed AI system meets stakeholders' and end-users' specific needs and expectations, as explicitly outlined in ISO/IEC 22989 (Information Technology — Artificial Intelligence — Concepts and Terminology). According to ISO standards, validation involves confirming through objective evidence that the requirements for a specific intended use or application have been fulfilled. In AI, validation goes beyond verifying

compliance with technical specifications—it assesses whether the system performs suitably in real-world conditions and scenarios.

The principal elements involved in the validation of edge AI systems encompass several dimensions:

- The identification of intended use and user requirements is foundational. Clear articulation and comprehensive understanding of user needs, operational contexts, and usage environments are paramount. This involves gathering input from stakeholders and end-users to form a robust basis for subsequent validation activities.
- Operational scenario definition is critical. Edge AI systems must be validated within scenarios that accurately represent real-world operational contexts. Scenarios are typically derived from realistic usage conditions, including normal operational states, boundary conditions, and potential abnormal or edge cases.
- Performance evaluation under realistic conditions is essential. Validation combines simulated environments and real-world testing to ensure edge AI systems perform reliably and effectively. Performance metrics, such as accuracy, precision, recall, robustness, resilience, and usability, form the basis for evaluating system performance and alignment with stakeholder expectations.
- Human-machine interaction and usability assessment are integral to validation. Edge AI systems are validated to ensure effective and intuitive interactions with human operators or users. Usability testing, user experience assessments, and feedback loops with real users facilitate comprehensive evaluations of the AI system's ease of use and accessibility.
- Safety, security, and ethical considerations are central elements of the validation process. These assessments verify that edge AI systems function correctly and comply with safety standards, security protocols, data privacy laws, and ethical guidelines, aligning with international frameworks and societal expectations.

The edge AI validation process typically involves structured, methodical steps:

- Requirement and Expectation Definition: Establishing clear validation criteria and user expectations, documenting them rigorously.
- Validation Planning: Creating detailed validation plans that specify methodologies, scenarios, test environments, and acceptance criteria.

- Scenario Development: Defining realistic operational scenarios and selecting representative use-cases and edge-cases for comprehensive validation.
- Simulation and Real-world Testing: Controlled simulations are conducted, followed by real-world trials to evaluate AI system performance against established criteria.
- Performance and Usability Assessment: Analysing performance outcomes, usability data, and user feedback to ascertain compliance with expectations and user requirements.
- Safety, Security, and Ethical Evaluation: Systematically reviewing compliance with safety and security standards, data protection requirements, and ethical norms.
- Reporting and Continuous Improvement: Compiling comprehensive validation reports, documenting findings and recommendations, and establishing iterative cycles for continuous system refinement.

Verification and validation of AI and edge AI models and data are required in safety-critical applications to ensure the trustworthiness of edge AI-enabled systems (e.g., reliability, availability, maintainability, safety, security, resilience, connectability, explainability, interpretability, transparency, etc.) as illustrated in Figure 1.3 and Figure 1.4.

Dependable edge AI systems involve using systems and software engineering principles to systematically guarantee dependability during the edge AI system's construction, V&V, and operation and consider legal and normative requirements directly from the start.

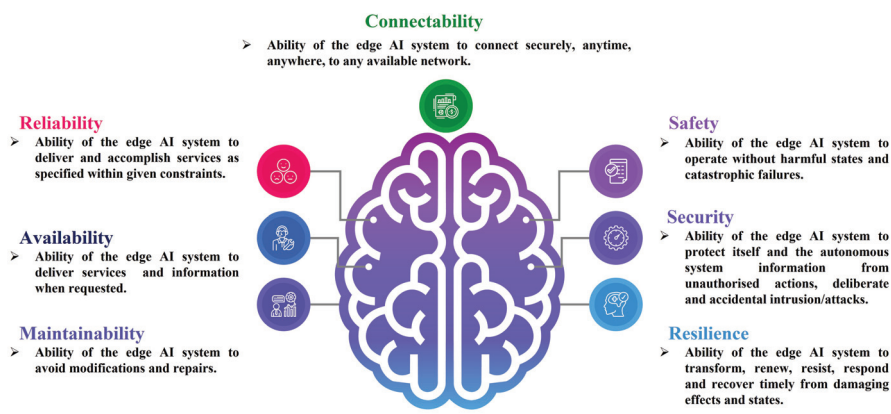

Figure 1.3 Edge AI dependability – Trustworthiness.

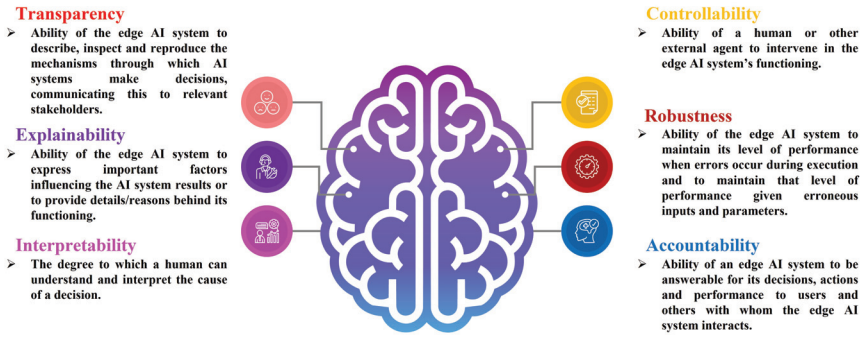

Figure 1.4 Edge AI dependability - Trustworthiness extended properties.

The progress made in developing standards and regulatory frameworks for AI and edge AI aims to ensure the responsible use of AI in various applications.

The relevant standards for AI that can be applied to edge AI systems are ISO/IEC 42001 and ISO/IEC TR 24028:2020 that are described below.

The ISO/IEC 42001 standard, a management system for AI, focuses on building trust and dependability in AI systems. It provides a framework to establish, implement, maintain, and continually improve their AI management systems, ensuring the responsible development and use of AI. The standard emphasises trustworthiness, fairness, transparency, and accountability in AI systems [43][44].

The ISO/IEC TR 24028:2020 standard addresses topics related to trustworthiness in AI systems, including approaches to establish trust in AI systems through transparency, explainability, controllability, etc.; engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security and privacy of AI systems [5].

Traditional V&V workflows, such as the V-model, are insufficient for ensuring the accuracy and reliability of AI and edge models. As a result, transformations of these workflows occurred to better serve edge AI applications.

1.2.1 Agentic AI and AI Agents

The evolution of generative AI and the emergence of AI agents and agentic AI requires addressing them under the presentation of foundational concepts

and edge AI verification and validation taxonomy by defining the concepts and their specific characteristics.

AI Agents can be defined as autonomous software entities engineered for goal-directed task execution within bounded digital environments. These agents are characterised by their ability to perceive structured or unstructured inputs, to reason over contextual information, and to initiate actions toward achieving specific objectives. The main characteristics of AI and edge AI agents are autonomy, task specificity, reactivity and adaptability, which enable the agents to operate as modular, lightweight interfaces between pre-trained AI models and domain-specific pipelines and workflows.

AI agents are the concrete instantiations of the agentic AI paradigm. An AI agent is a specific software or hardware entity that embodies the principles of agentic AI. It is a tangible system equipped with sensors to perceive its environment and effectors to act upon it. While agentic AI is the "what," the AI agent is the "how", the actual implementation that performs tasks, makes decisions, and interacts with external environments.

Agentic AI systems describe a paradigm shift from isolated AI agents to collaborative, multi-agent ecosystems capable of decomposing and executing complex goals [21]. These systems typically consist of orchestrated or communicating agents that interact via tools, APIs, and shared environments [23][14].

A key distinction between agentic AI and AI agents lies in their level of abstraction, as the agentic AI is a conceptual framework, whereas an AI agent is a functional system. An analogy can be drawn between the theory of computation and a physical computer. One provides the theoretical foundation and a model of what is possible, while the other is the practical machine that executes computations based on that theory.

Agentic AI reflects a broad paradigm in AI and edge AI centred on creating systems that can perceive their environment, reason about their observations, and act autonomously to achieve specific goals. It is the underlying philosophy and set of principles that guide the development of intelligent, goal-oriented systems. This concept emphasises proactivity, reactivity, and social ability, defining the potential for AI to operate as an independent actor rather than a passive tool. Agentic AI systems introduce internal orchestration mechanisms and multi-agent collaboration frameworks. Agentic AI extends the foundational architecture to support complex, distributed, and adaptive behaviours by integrating components such as specialised agents, persistent memory, orchestration and advanced reasoning and planning. Agentic AI introduces novel memory integration, communication

paradigms, and decentralised control, paving the way for the next generation of adaptive workflow automation in autonomous systems, swarm robotics, and autonomous vehicles with scalable, adaptive intelligence.

In robotics and automation, agentic AI enables collaborative behaviour in multi-robot systems. Each robot operates as a task-specialised agent, such as a picker, transporter, or mapper, while an orchestrator supervises and adapts workflows. These architectures rely on shared spatial memory, real-time sensor fusion, and inter-agent synchronisation for coordinated physical actions. Use cases include warehouse automation, drone-based orchard inspection, and robotic harvesting [25].

Verification and validation of edge AI systems, based on AI agents and agentic AI components, must focus on ensuring the correctness, reliability, and robustness of autonomous decision-making in highly dynamic and constrained environments, which requires validating that the AI agents consistently perform their intended functions correctly under varying external environment conditions, including unexpected scenarios and adversarial inputs.

Due to the limited computational resources typical of edge devices, V&V must also confirm that the system meets stringent real-time performance requirements, ensuring timely responses to critical events despite hardware and network limitations.

Another aspect is assessing the resilience and safety of adaptive learning processes within these systems, particularly as they evolve in open environments. V&V efforts should capture how individual agents and collective multi-agent behaviours emerge and interact, verifying alignment with overall system objectives and preventing unsafe or unintended actions.

Additionally, transparency and trustworthiness are key elements that enable human oversight, offering clear traceability and checkability of decisions made by autonomous components at the edge.

1.3 Defining Verification and Validation per Standard

Several ISO standards offer consistent definitions for verification and validation, primarily within the context of quality management and systems/software engineering that can be applicable to AI and edge AI as presented below.

ISO 9000:2015 (Quality management systems - Fundamentals and vocabulary) provides the definition for verification as the “confirmation,

through the provision of objective evidence, that specified requirements have been fulfilled" [29]. The focus is on confirming that the system or component conforms to its design specifications and requirements [31]. It answers the question: “Did we build the product right?” [32]. Verification is often viewed as an internal process comparing the outputs of a development phase against the inputs [31]. Validation is defined as the “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled” [29]. The focus shifts to ensuring the system meets the needs of the user and fulfils its intended purpose in the actual context of use. It answers the question: “Did we build the right product?” [32]. Validation often involves testing under real or simulated use conditions and considers stakeholder needs [31].

ISO 9001:2015 (Quality management systems - Requirements), states that validation activities ensure the resulting products/services meet requirements for the specified application or intended use. Validation often involves acceptance testing with end-users and assessing fitness for purpose, making it frequently an external process, whereas verification is more often internal. Both verification and validation are essential components of quality management and are necessary for ensuring a dependable system [30].

ISO/IEC/IEEE 15288:2015 (Systems and software engineering - System life cycle processes) standard integrates V&V into the system lifecycle and considers that the **verification process** has as purpose “to provide objective evidence that a system or system element fulfils its specified requirements and characteristics” [34]. It involves activities comparing the system or element against requirements, design descriptions, and other required characteristics, confirming it was “built right” [35], while the **validation process** has as purpose “to provide objective evidence that the system, when in use, fulfils its business or mission objectives and stakeholder requirements, achieving its intended use in its intended operational environment” [33]. This process confirms that stakeholder requirements are correctly defined, and that the system meets its intended purpose in the context where it will operate [33].

ISO/IEC 22989:2022 (Information technology - Artificial intelligence - Artificial intelligence concepts and terminology) AI-specific standard defines **verification** as “confirmation, through the provision of objective evidence, that specified requirements have been fulfilled,” noting it assures conformance to specification [9]. While not explicitly defining validation in the same way, it defines **trustworthiness** as the “ability to meet stakeholder

expectations in a verifiable way" [38]. This definition links the core goal of validation (meeting stakeholder expectations/needs) directly to the concept of trustworthiness in AI. The standard also incorporates a "verification and validation" phase within its depiction of the AI system lifecycle [39].

ISO/IEC TR 24028:2020 (Information technology - Artificial intelligence - Overview of trustworthiness in Artificial Intelligence) technical report further reinforces the link between validation and trustworthiness and defines **trustworthiness** as the "ability to meet stakeholder expectations in a verifiable way" [40]. This aligns the concept of trustworthiness directly with the objective of validation – confirming that stakeholder needs and intended use requirements are met [42]. The report discusses assessing and achieving key characteristics like reliability, safety, security, and privacy, all crucial aspects evaluated during validation [41].

ISO/IEC 42001:2023 (**AI Management System)** standard specifies requirements for establishing, implementing, maintaining, and continually improving an AI Management System (AIMS) within an organization [43]. An AIMS provides a structured framework for responsible AI governance, risk management, and operational control throughout the AI lifecycle [44]. Verification activities are integral to an AIMS, supporting risk assessment, impact assessment, performance evaluation, and ensuring compliance with policies and objectives [44]. Notably, ISO/IEC 22989 (providing the core AI terminology) is a normative reference for ISO/IEC 42001, highlighting the foundational role of clear definitions [45].

IEEE 1012-2016 (IEEE Standard for System, Software, and Hardware Verification and Validation) standard applies to systems, software, and hardware being developed, maintained, or reused (legacy, commercial off-the-shelf [COTS], non-developmental items) [91]. The term "software" also includes firmware and microcode. Additionally, each of the terms "system," "software," and "hardware" encompasses documentation. V&V processes include the analysis, evaluation, review, inspection, assessment, and testing of products. V&V processes are used to determine whether the development products of a given activity conform to the requirements of that activity and whether the product satisfies its intended use and user needs. V&V lifecycle process requirements are specified for different integrity levels. The scope of V&V processes encompasses systems, software, hardware, and their interfaces.

1.4 Key Elements for Edge AI Verification and Validation

The elements for verifying and validating edge AI may encompass operational aspects, system integration, AI models and human-machine interaction. Verification and validation are important to ensure reliability, performance and accuracy of complex systems.

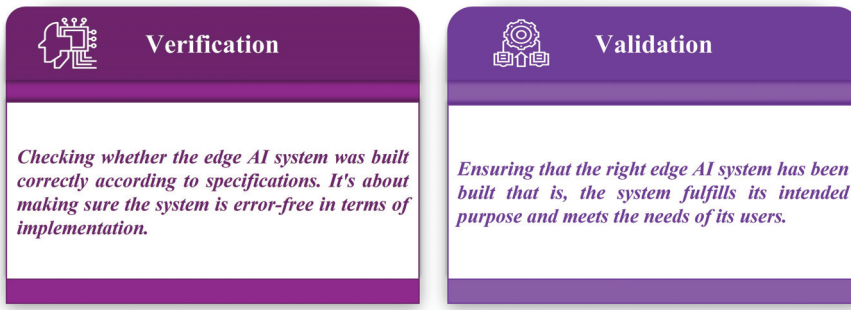

Figure 1.5 Verification and validation.

Edge AI system verification and validation refers to the processes and methodologies used to ensure that an edge AI system is dependable, performs as expected, and meets certain standards before it is deployed.

These processes are crucial as the edge AI algorithms can work with high-stakes decision-making, various sizes datasets, learn and evolve over time. The processes are needed for ensuring that the edge AI systems do what they are supposed to do, without unintended consequences, biases, or errors.

AI and edge AI systems typically focus on the actual algorithms and models to ensure that they perform as intended under various conditions. In addition, edge AI systems focus on validating the systems performance on resource-constrained devices, network conditions and privacy in real-world scenarios.

It is critical to distinguish verification from validation. While verification checks conformance to specifications ("Did we build the system right?"), validation confirms that the system meets the needs of the customer and other stakeholders and fulfils its intended purpose in its operational environment ("Did we build the right system?") [30].

The introduction of AI in product and systems development has significantly increased the complexity of electronic components and systems (ECS), by integrating various technologies such as hardware, software, ML,

DL, NNs, generative AI, and advanced data analytics. This complexity necessitates robust verification and validation frameworks and benchmarking to ensure these systems operate correctly and efficiently as illustrated in Figure 1.6. Complex edge AI models require verification and validation to ensure their predictions, decisions, and content generation outputs are reliable and accurate, which is critical for maintaining the trustworthiness of AI systems. Failures in edge AI-based ECS can have significant economic and business-critical consequences, including system failures, financial loss, and damage to infrastructure, making the dependability of edge AI systems paramount.

In machine vision, specific verification concerns arise from the need to ensure reliable object detection, tracking, segmentation, or pose estimation across a wide range of dynamic conditions. For example, verification must confirm that visual inference results remain stable under varying lighting, occlusion, and motion blur, common challenges in edge deployments like factory floors or drones.

Ensuring robustness and reproducibility in edge-based machine vision systems is inherently difficult due to the high variability and noise in visual data. Unlike structured tabular inputs, images and videos exhibit a vast range of intra-class variation—objects or actions belonging to the same class can appear drastically different depending on factors such as:

- Lighting conditions (e.g., shadows, reflections).
- Occlusions or partial views.
- Background clutter.
- Camera distortions, blur, or motion artifacts.
- Variability in object shape, colour, texture, or viewpoint.

A comprehensive V&V framework, presented in Figure 1.6, along with benchmarking of edge AI-based methods, frameworks, tools, and ECS, is essential to ensure performance and dependable system properties like security, reliability, robustness, and fairness.

Verification ensures that edge AI-based methods, frameworks, tools, and electronic components and systems are built correctly and meet specifications, while validation confirms they perform as intended in real-world scenarios.

In edge AI systems there is a need of creating a structured approach to defining and applying such a framework to edge AI-based tools and methods, ensuring ECS meet functional and non-functional requirements, quality, KPIs, and performance standards.

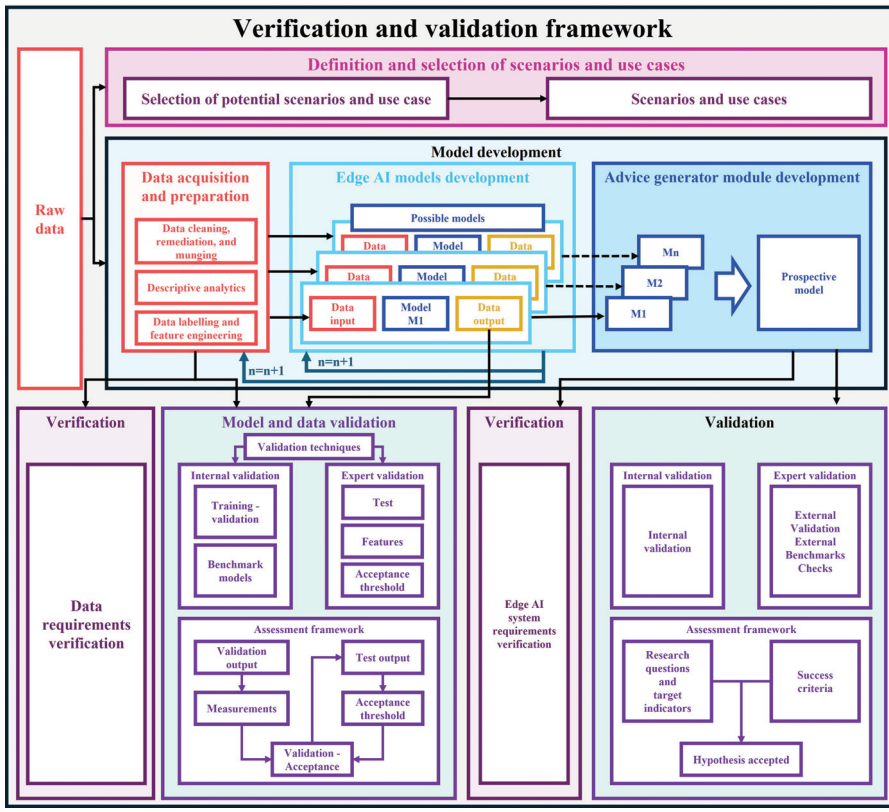

Figure 1.6 Verification and validation framework.

1.4.1 Core Elements for AI Verification

Verification activities in AI systems must address multiple facets, spanning data, models, system-level behaviour, and the processes governing development and deployment. Ensuring the integrity and appropriateness of each element is crucial for overall system trustworthiness.

1.4.1.1 Data Verification

Given that many AI and edge AI systems, particularly those based on ML, learn from data, verifying the data itself is paramount [3]. Key aspects include:

Data Quality: Assessing if the data meets predefined standards for accuracy, completeness, consistency, timeliness, and representativeness for the

target domain [3]. This involves checking for errors, missing values, correct formatting, and ensuring the data is current and relevant [46]. Poor data quality directly impacts model performance and reliability. Machine vision applications often require extensive data augmentation and synthetic dataset generation for robustness. In this case, the validity of the augmented dataset needs to be verified for plausibility and compliance with conditions of the actual use of the system.

Data Bias: Identifying systemic skews or prejudices within the data that could lead to unfair or discriminatory outcomes [3]. Verification involves confirming that bias detection methods have been applied and that any mitigation steps align with fairness requirements or definitions. This includes checking for underrepresentation or imbalances across demographic groups [3].

Data Provenance and Lineage: Ensuring the origin and history of the data are understood and documented, including all transformations and processing steps [47]. Verification confirms traceability back to authorized sources and validates the integrity of the data pipeline.

Data Security and Privacy: Confirming that data collection, storage, and processing adhere to relevant privacy regulations like General Data Protection Regulation (GDPR), a law in the European Union aimed at safeguarding the data and privacy of EU residents or California Consumer Privacy Act (CCPA) a US state law that applies to for-profit businesses operating in California that collect personal information from California residents, and organizational security policies [3]. This includes verifying the implementation of techniques like anonymization, encryption, access controls, and proper consent management [48].

Data Labelling: For supervised learning, verifying the accuracy, consistency, and quality of labels applied to the training and testing data is crucial, as errors here directly impact model learning [3].

1.4.1.2 Model Verification

The AI and edge AI model itself, the core component that performs learning and prediction, requires rigorous verification:

Accuracy and Performance: Quantifying how well the model achieves its intended task according to predefined metrics (e.g., precision, recall, F1-score for classification; BLEU score for translation) evaluated on unseen test or validation datasets [41]. Verification confirms that the achieved performance meets the specified requirements or benchmarks.

Robustness: Evaluating the model's ability to maintain its performance level when faced with noisy data, adversarial perturbations, changes in data distribution (drift), or other unexpected conditions [49]. Verification checks if the model's resilience meets specified criteria under defined stress conditions.In machine vision models, robustness testing should also include tests for perceptual artifacts, such as camera motion blur or lens distortion, and adversarial perturbations that affect visual features. This is especially important for safety-critical applications like automated visual inspection or autonomous guidance.

Reliability: Assessing the consistency and predictability of the model's outputs under normal operating conditions over time [50]. Verification aims to confirm that the model behaves dependably within its specified operational domain.

Efficiency: Measuring the model's consumption of computational resources, such as processing time, memory usage, and energy [48]. Verification ensures the model operates within the constraints imposed by the deployment hardware or system requirements.

1.4.1.3 System-Level Verification

Verification must also extend to the AI and edge AI system, considering its interaction with its environment and users:

Safety: Confirming that the system operates without causing unacceptable levels of risk or harm to humans, property, or the environment [49]. This involves verifying adherence to specific safety requirements, standards (like ISO 26262 for automotive), and risk assessments. In vision-driven systems, safety verification must ensure that the interpretation of the visual scene cannot trigger unsafe behaviour due to false positives or misclassifications e.g., mis detecting a pedestrian or failing to recognise a hazard in the camera feed.

Security and Resilience: Checking the implementation and effectiveness of measures designed to protect the system against threats like unauthorized access, data breaches, model tampering, and adversarial attacks [3]. It also includes verifying the system's ability to withstand and recover from disruptions [51].

Fairness: Evaluating system outcomes across different demographic or user groups to ensure equity and the absence of harmful bias or discrimination, according to defined fairness metrics or criteria [3].

Privacy: Verifying that the system's operation, including data handling and output generation, complies with privacy principles and regulations throughout its use [52].

1.4.1.4 Process and Governance Verification

Beyond the technical components, the processes surrounding the AI system also require verification of:

Transparency: Assessing whether sufficient and appropriate information about the AI system (its purpose, data sources, model type, limitations, performance) is documented and made available to relevant stakeholders (developers, deployers, users, regulators) [53]. Verification checks if documentation and communication channels meet specified transparency requirements.

Explainability and Interpretability: Evaluating whether the system can provide understandable reasons or justifications for its outputs or decisions, tailored to the applications and users. Verification checks if the explanation mechanisms provided meet requirements for clarity, fidelity, and utility.

Accountability: Confirming that clear roles, responsibilities, governance structures, and mechanisms for oversight, audit, and redress are defined, documented, and effectively implemented [6]. Verification involves auditing these governance processes and structures against standards like ISO/IEC 42001.

These verification elements are deeply interconnected [3]. For instance, verifying fairness requires access to appropriate data and potentially explainability techniques to understand model behaviour. Verifying safety may depend on demonstrating model robustness and having transparent documentation of system limitations. An opaque model hinders the verification of its internal logic, making it difficult to assess its safety or fairness properties directly. This interdependence necessitates a holistic verification strategy rather than treating each element in isolation.

Furthermore, the emphasis placed on different verification elements naturally shifts depending on the type of AI system. For data-driven ML models, verification heavily scrutinizes data quality, bias, model performance, and robustness [3].

In contrast, for symbolic AI systems built on explicit rules and logic, verification may concentrate more on the consistency, correctness, and completeness of the knowledge base and the soundness of the reasoning engine [64].

Hybrid neuro-symbolic systems demand verification of both the neural and symbolic parts, as well as their complex interactions, representing a distinct verification challenge [64].

1.4.2 Core Elements Subject to AI Validation

Given validation's focus on fitness for purpose and meeting stakeholder needs, the elements assessed extend beyond traditional software checks. Validating AI systems requires evaluating a broader spectrum of characteristics that reflect their performance, usability, effectiveness, and impact within their socio-technical context [6]. The exact scope may vary based on the application domain, but the core elements subject to validation include:

1.4.2.1 Ensuring Fitness for Intended Purpose and Operational Context

This is a central element of validation. It involves confirming that the AI system effectively achieves its stated goals within the specific environment and conditions of its intended use [54]. This requires a clear definition of the intended purpose and the Operational Design Domain (ODD), the specific conditions under which the system is designed to function. However, defining and validating against these can be particularly challenging for adaptive AI systems or those designed for open-world environments where conditions are dynamic and unpredictable [3]. Validation must assess performance not just under nominal conditions but also under stress, edge cases, and potential environmental shifts or adversarial inputs [85]. Frameworks like the NIST AI Risk Management Framework (RMF) emphasize establishing context (Map function) as a foundational activity to inform subsequent measurement and management, including validation [55].

1.4.2.2 Meeting User Needs and Stakeholder Expectations

Validation explicitly confirms that the system satisfies the requirements and expectations of its end-users and other relevant stakeholders [30]. This extends beyond purely functional requirements to encompass aspects like usability, user satisfaction, ease of integration into existing workflows, and alignment with business objectives [56]. Because AI systems can impact a wide range of individuals and groups, validation should involve engagement with diverse stakeholders, including end-users, domain experts, potentially affected communities, and regulators, to capture a comprehensive set of needs and expectations [90]. Addressing the challenge that these needs

might be implicit, diverse, or even conflicting is a key part of the validation process [57].

1.4.2.3 Assessing Real-World Effectiveness and Outcomes

Validation must measure how the AI and edge AI system performs in practice, assessing its actual effectiveness in achieving desired outcomes within realistic scenarios [59]. This moves beyond performance metrics derived solely from laboratory settings or curated test datasets. It involves evaluating the system's impact on relevant Key Performance Indicators (KPIs), operational efficiency, safety records, cost savings, or other context-specific measures of success [58]. Initiatives like NIST's Assessing Risks and Impacts of AI (ARIA) program are specifically focused on developing methodologies to measure these real-world impacts under controlled conditions [61]. This assessment typically requires methods such as Operational Testing (OT), field testing, pilot deployments, and continuous performance monitoring after deployment [6]. In edge machine vision systems, this includes validating that visual perception models continue to perform accurately when deployed with quantized weights, compressed inputs, or on hardware that introduces latency jitter. This real-world validation should account for degradation due to environmental variables and resource limitations.

1.4.2.4 Evaluating Usability and Human-AI Interaction

For edge AI and AI systems that interact with or support humans, validation must assess the quality and effectiveness of this interaction [60]. This includes evaluating usability (ease of use, learnability, efficiency), the clarity and utility of the interface, the cognitive load imposed on the user, and overall user satisfaction [63]. Particularly for human-AI collaboration or teaming scenarios, validation needs to assess the effectiveness of the partnership, the safety of the interaction, the appropriateness of trust levels (avoiding over-trust or under-trust), and the degree of shared understanding between human and AI [61]. This requires human-centered evaluation methods, such as usability studies, task analyses involving representative users, and systematic collection of user feedback [62].

1.4.2.5 Validating Ethical Alignment and Societal Impact

A critical dimension of edge AI validation involves assessing the system's alignment with ethical principles and societal values [6]. This includes validating characteristics like fairness, accountability, and transparency in practice [55]. Methodologies such as Ethical Impact Assessments (EIAs) are

emerging to help proactively identify, assess, and mitigate potential negative ethical and societal consequences before and during deployment [61]. A key focus is validating fairness and non-discrimination, moving beyond simple dataset metrics to assess the actual impact on different demographic groups in real-world deployment contexts [90]. This also involves considering broader societal implications related to employment, environmental sustainability, and the functioning of democratic processes [7].

1.4.2.6 Data Quality and Suitability

High-quality data ensures that models are trained effectively and can make accurate predictions in real-world scenarios [36]. As AI and edge AI systems become more complex and are deployed in diverse environments, the challenges associated with data quality and suitability have become increasingly significant. Considering the specific requirements for various AI and edge AI systems, challenges for data quality and suitability in AI and edge AI validation include:

Relevance and Representativeness: the data used for training and validation is relevant and representative of the real-world environment in which the AI and edge AI systems operate. Data must reflect the diversity of conditions, contexts, and populations that the system will encounter. If the training data is biased or unrepresentative, the model's performance may deteriorate when applied to actual situations.

Volume and Availability: Considered very important, especially in scenarios where data may be generated at high velocity. Obtaining enough high-quality data for training and validation can be difficult. In many cases, developers may struggle to gather sufficient diverse data from edge devices, leading to models that are not well-trained for all possible situations they may encounter in deployment.

Label Quality: important for supervised learning, as it directly impacts model accuracy. Inaccurate or inconsistent labelling can mislead the training process and result in poor performance in operational environments. Ensuring the reliability of labels, especially when data is labelled manually or derived from semi-automated processes, can be a significant extra work.

Bias and Fairness: the biases in learning and training of data, can lead to outcomes that are unfair when models are deployed. AI and edge AI systems trained on biased data may perpetuate existing stereotypes or discriminate against certain classes and groups. Addressing data bias and ensuring fairness

in model predictions is key to building trustworthy AI and edge systems that serve all stakeholders equitably.

Data Drift: refers to the shifts in data distributions over time, which can degrade model performance. As the underlying data evolves, models may become less accurate or irrelevant. Ongoing monitoring and adaptation of models are necessary to mitigate the effects of data drift, making it a continuous challenge for AI and edge AI validation.

Data Preprocessing: is a critical step in data management, particularly for edge AI systems with limited resources. Cleaning and transforming data into suitable formats can be challenging, when using with diverse data sources and formats. This preprocessing must be efficient to ensure real-time performance while maintaining data accuracy and integrity.

Synthetic Data: helps augment training datasets and has several limitations. The effectiveness of synthetic data depends on its ability to mimic real-world scenarios accurately. If synthetic data does not accurately represent the complexities of real-world environments, it may lead to models that underperform when applied to actual data.

Edge-Specific Challenges: these are related to data collection from distributed edge devices, considering elements like latency, bandwidth constraints, and intermittent connectivity, which can complicate the data validation process. Ensuring data quality in these scenarios requires innovative approaches to data management and model training.

1.5 The Edge AI Verification and Validation Lifecycle

According to the OECD recommendation on artificial intelligence [10], an AI system is a machine-based framework that, driven by either explicit or implicit goals, deduces from the input it receives how to produce outputs such as predictions, content, recommendations, or decisions that may impact physical or virtual environments. The levels of autonomy and adaptability of different AI systems can vary after they are deployed. The lifecycle of an AI system generally encompasses multiple stages, which include planning and design; data collection and processing; model development and/or adaptation of existing models for specific tasks; testing, evaluation, verification, and validation; deployment for use; operation and monitoring; and retirement or decommissioning. These stages often occur iteratively and are not strictly

linear. The choice to retire an AI system can be made at any time during the operation and monitoring stage.

AI and edge AI systems are distinct from other system types, which can influence the processes of the lifecycle model, such as [9]:

- Most SW systems are designed to operate in exactly defined manners dictated by their requirements and specifications. In contrast, AI and edge AI systems that utilize ML rely on data-driven training and optimization techniques to address a wide range of inputs.
- Traditional SW applications tend to be predictable, whereas this is less frequently true for AI and edge AI systems.
- Additionally, traditional SW applications are generally verifiable, while evaluating the performance of AI and edge AI systems often necessitates statistical methods, making their verification more complex.
- AI and edge AI systems usually require numerous iterations of enhancement to reach satisfactory performance levels.

The edge AI development lifecycle outlines the stages involved in creating and operationalizing edge AI systems. It starts with problem definition, functional, non-functional requirements and data collection, followed by data preparation and feature engineering. Model selection and architecture design precede the training phase, where algorithms learn from the prepared dataset. Validation and testing ensure model performance and generalization. Iterative refinement optimizes the model based on results. Deployment integrates the AI system into production environments. Monitoring and maintenance track performance, address drift, and update the model as needed.

Embedding AI and generative AI into the system design requires shifting from the current V-model, which addresses the HW and SW development cycle, to a W-model superimposed on the V-model to account for data and AI-specific artifacts. This includes the AI model development and data into the AI system's lifecycle development, as illustrated in Figure 1.7 [92].

When superimposed on the traditional V-model used in HW and SW development, the W-model AI development lifecycle creates a comprehensive framework that addresses the distinct yet interrelated processes of AI data development and HW/SW engineering. This approach ensures that AI models and supporting systems are developed in a cohesive, iterative, and validated manner. This approach aligns existing tools and methods with AI technologies. The extension into the W-model structures represents the development workflow of AI systems comprising HW, SW, AI stack, and data components.

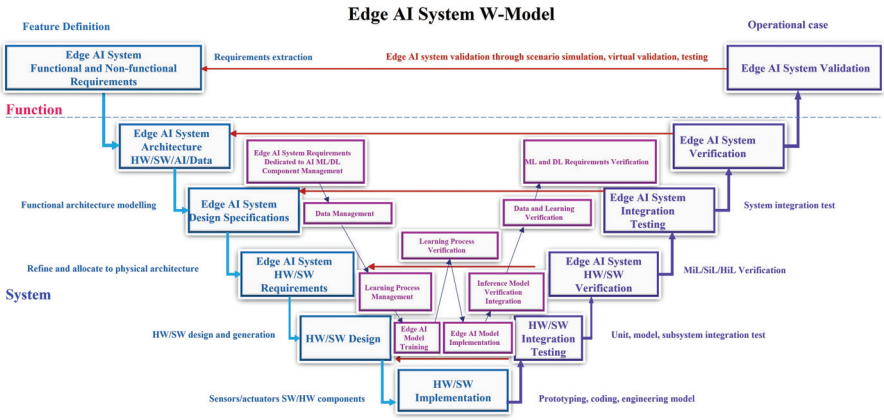

Figure 1.7 Edge AI system W-Model (adapted from [92])

The inner W part of the model represents the AI-enabled processes and workflows integrated into the conventional model. The AI system W-model emphasises systematic validation and verification at each stage of the AI development process, helping ensure the robustness, reliability, and performance of AI systems. The W-model addresses the specific design and development requirements of edge AI systems, distinguishing them from traditional HW/SW and computing paradigms. The novelty in the model lies in the fact that the data required for development and AI, ML/DL, and generative AI model training is integrated into the development cycle, superimposed on the traditional V-model, and follows the algorithm selection and training in each lifecycle stage.

The edge AI system W-model emphasises that AI and generative AI are integral to the lifecycle development processes of any AI-based product or service.

As presented in the AI system W-model at the start of the development lifecycle, developers can utilise AI and generative AI to understand domain requirements and design architecture. The design captures both functional and non-functional requirements for embedded computing systems, such as those in automotive control or industrial units, considering hardware constraints and real-time performance needs. Challenges concerning edge AI requirements and AI requirements engineering are extensive and due in part to the practice by some to treat the AI element as a "black box". Formal specification has been attempted and has proven to be difficult for tasks that are hard to formalise, requiring decisions on the use of quantitative

or Boolean specifications, as well as the incorporation of data and formal requirements. The challenge is to design effective methods for specifying both desired and undesired properties of systems that utilise AI- or ML-based components.

When considering the broader principles of agentic AI and AI agents, their development must be integrated into the edge AI system W-model as a specific part of the lifecycle development processes of any AI-based product or service. As a result, the agentic AI and AI agents V&V extends beyond the individual agent, focusing on validating the system's autonomy and the ability to achieve long-term goals without unexpected consequences by assessing the alignment of the AI agent's goals with the overall objectives of the edge AI system and ensuring that its learning and adaptation mechanisms do not lead to unsafe or undesirable states over time.

1.6 Failure Case Behaviour in Edge-based Machine Vision Systems

In the context of edge-based machine vision systems, the study of failure case behaviour is a critical component of any robust verification and validation framework. These systems are increasingly deployed in real-world, safety-critical environments, ranging from autonomous vehicles and industrial robotics to surveillance and medical diagnostics, where failures can result in substantial consequences. While conventional validation focuses on average-case performance metrics such as accuracy or mean average precision, these metrics often obscure rare but consequential failure modes. An edge AI system that performs well under ideal conditions may fail unexpectedly in the presence of visual distortions, environmental variability, or edge hardware constraints.

Failures in machine vision models frequently arise in conditions that deviate from the data distribution seen during training. Examples include poor lighting, motion blur, occlusions, scale variation, or visual clutter. In edge deployments, such conditions are not only likely but expected, and the consequences of misclassification or missed detection can be severe. Furthermore, edge systems often operate with limited fallback options, and they must respond in real time, leaving little margin for error recovery. Understanding and characterizing these failure scenarios is therefore essential for both safety assurance and iterative model improvement.

One important strategy for investigating failure modes involves deliberate stress testing through visual perturbations. By applying controlled

transformations—such as adding noise, blurring, shifting brightness, or introducing occlusions—it becomes possible to evaluate how resilient a vision model is to real-world distortions. These tests often reveal brittle model behaviours that are not apparent during standard validation.

In addition, targeted scenario-based testing using simulation tools or recorded video sequences enables systematic exploration of edge cases. This is especially valuable for applications involving dynamic environments, such as autonomous navigation, where rare events (e.g., unexpected pedestrian appearance or sensor occlusion) may not be captured adequately in available datasets. Scenario replay or simulation also supports reproducibility of observed failures, which is often a challenge in field deployments.

Another aspect of failure case analysis is the examination of uncertainty and confidence levels in model predictions. Machine vision systems may produce incorrect predictions with unjustified confidence, especially when faced with unfamiliar or out-of-distribution inputs. Monitoring softmax confidence, prediction entropy, or Bayesian uncertainty estimates can help identify instances where the model is likely to fail. These signals may be used in runtime monitoring or to trigger fail-safe mechanisms.

Post-hoc explainability methods, such as saliency maps or activation heatmaps, also play an important role in understanding failure behaviour. By visualising the regions of an input image that contributed most to a model's prediction, one can diagnose whether a failure was due to the model focusing on irrelevant or misleading features. This insight often reveals underlying dataset biases or spurious correlations that were inadvertently learned. By combining scenario condition variables and the predictions as features in probabilistic frameworks (e.g., Bayesian networks) is also a method for modelling the uncertainty in predictions.

Hardware-specific issues also need to be considered in failure analysis. For instance, the quantization of weights and activations required for execution on edge hardware (e.g., FPGAs or ASICs) can introduce numerical inaccuracies that degrade model performance in subtle ways. Testing the consistency between floating-point reference models and their hardware-deployed counterparts is essential to identify precision-induced errors. Similarly, real-time system profiling can reveal frame drops, synchronization mismatches, or input-output latency violations that lead to perceptual failures.

Finally, insights gained from the analysis of failure cases should feed back into the design and development process. Difficult or misclassified examples can be incorporated into retraining pipelines, synthetic data can be generated

to increase robustness, and system architectures can be adapted to detect and respond to high-uncertainty inputs. Ideally, safety envelopes are defined at design time to formally capture the operational conditions under which the system is guaranteed to function correctly. This creates a closed-loop process that not only identifies but also mitigates and prevents known failure patterns.

The systematic study of failure case behaviour is indispensable for building trustworthy machine vision systems on edge platforms. It enables developers to move beyond average-case performance toward comprehensive assurance of correctness, robustness, and safety under realistic and adverse conditions.

1.7 Research Challenges in Edge AI Verification and Validation

Edge AI represents a cutting-edge computing approach that seeks to relocate the training and inference of ML models to the network's edge [12]. However, implementing intelligence at the edge presents several significant challenges, such as the necessity to limit model architecture designs, ensuring the secure distribution and execution of the trained models, and managing the considerable network load needed to disseminate the models and the data gathered for training.

Edge AI systems that incorporate continuous learning involve the gradual updating of the models within the systems during production and test runs operations [9]. The data input into a system during these operations, is not only evaluated to generate an output but is also concurrently utilized to modify the model, aiming to enhance it based on the production data. Depending on the design of the continuous learning system, certain human interventions may be necessary, such as data labelling, validating the application of specific incremental updates, or monitoring the performance of the edge AI system. Continuous learning can address the limitations of the initial training data and assist in managing data drift and concept drift, but it also presents significant challenges in ensuring the edge AI system operates correctly while learning. It is essential to verify the system in production and to capture the production data to be able to include them as part of the training dataset in future system updates.

Catastrophic interference (and catastrophic unlearning) occurs when the training for new tasks disrupts the model's comprehension of previous tasks [11]. As new information supersedes earlier learning, the model forfeits its

capability to manage its initial tasks. Given the risk of catastrophic interference, continuous learning necessitates the capability to learn over time by integrating new observations from current data while preserving prior knowledge [9]. Numerous ML algorithms excel at learning tasks only when the data is provided in a single batch. As a model is trained on a specific task, its parameters are modified to effectively tackle that task. However, when new training data is introduced, the adjustments made for these new inputs can erase the knowledge the model had previously gained. In the context of neural networks, this occurrence is regarded as one of their key limitations.

The combination of edge AI, IoT and Cyber-Physical Systems (CPS) marks a significant transformation in data processing by bringing it closer to the origin. This strategy minimizes latency, improves real-time decision-making, and lessens the load on centralized cloud resources. In CPS, control logic is utilized to process input from sensors, through actions of actuators and thus affecting processes occurring in the physical world [9]. This is particularly evident in robotics, where sensor data is directly employed to manage the robot's operations and execute tasks in the physical world. Typically, robots are equipped with sensors at the edge to evaluate their current conditions, processors to facilitate control through analysis and action planning, and actuators to implement those actions.

In contrast to industrial robots, which are consistently repeating the same trajectories and actions without deviations, service robots or collaborative robots must adapt to evolving situations and dynamic environments [9]. Programming this adaptability presents significant challenges due to the inherent variability. Components of edge AI systems can play a role in the control software and planning processes through the "Sense-Plan-Act" framework, allowing robots to modify their actions in response to obstacles or changes in the location of target objects. The integration of robotics and edge AI system components facilitates automated physical interactions with objects, environments, and individuals.

In machine vision-based robotics, the visual processing pipeline itself must be verified and validated not only for accuracy but also for real-time responsiveness. Edge V&V must ensure that latency from image acquisition to action initiation does not exceed application-specific safety thresholds. Techniques like real-time trace logging and FPGA-based image path profiling can support this validation.

The challenges and appropriate methodologies for AI verification are not uniform; they vary significantly depending on the type of AI model employed and the application domain's risk profile.

Model-Specific Challenges and Verification Focus

Deep Learning (DL) / Sub-Symbolic AI:

- *Challenges:* The primary verification challenges stem from their inherent opacity (making internal logic inscrutable) [64], strong data dependency (performance tied to training data quality and representativeness) [76], difficulty in formal specification of complex learned behaviours [64], susceptibility to adversarial examples, and challenges in generalization beyond training data distributions [76]. Scalability of verification methods is a major bottleneck due to the vast number of parameters and high-dimensional inputs [52]. Non-determinism can also arise during training or inference [70].
- *Verification Focus:* Emphasis is placed on empirical performance evaluation using diverse test datasets, extensive robustness testing against perturbations and adversarial attacks, fairness audits to detect biases learned from data, applying explainable AI (XAI) techniques (like LIME, SHAP, saliency maps) and interpretable AI (IAI) to gain insights into model decisions [74][75] and, where feasible, formal verification of specific, localised properties such as robustness bounds around specific inputs [69].

Symbolic AI / Rule-Based Systems:

- *Challenges:* These systems often suffer from **brittleness**, meaning they struggle to handle situations not explicitly covered by their predefined rules or knowledge base [73]. Creating and maintaining large, consistent, and complete knowledge bases can be labour-intensive and requires significant domain expertise [72]. They typically lack the ability to learn directly from raw, unstructured data. A computer vision model trained to detect stop signs may misclassify a slightly occluded or weathered sign because it hasn't seen enough variation in training. Traditional software can also exhibit brittleness, i.e. they both struggle but in different forms.
- *Verification Focus:* Verification centres on the logical integrity of the system. This includes checking the consistency of the rule set and knowledge base (absence of contradictions), analysing completeness (do the rules cover the intended domain?), formally verifying logical properties like soundness and validity of reasoning steps [64] and ensuring the traceability of outputs back to specific rules, which provides inherent explainability [72].

Neuro-symbolic AI:

- *Challenges:* This hybrid approach aims to combine the strengths of DL and symbolic AI but verifying the interaction and ensuring consistency between the neural (learning) and symbolic (reasoning) components is a key challenge [64]. Developing unified V&V frameworks that can handle both paradigms simultaneously is an active area of research [64].
- *Verification Focus:* Requires a multi-pronged approach: verifying the neural components using DL-specific techniques, verifying the symbolic components using logic-based methods, and crucially, verifying the interface and the correctness of the combined system's behaviour. A major research direction involves leveraging the symbolic part to constrain, explain, or formally verify aspects of the neural part's behaviour [64].

Domain-Specific Challenges and Verification Focus

Safety-Critical Systems (e.g., Automotive, Aerospace, Medical, Industrial Control):

- *Requirements:* These domains demand high levels of reliability, safety, robustness, and predictability [49]. System failures can have catastrophic consequences, including loss of life, severe injury, or significant environmental damage [49].
- *Challenges:* The need for provable guarantees clashes with the opacity and non-determinism of many AI components [65]. Meeting stringent regulatory standards (e.g., ISO 26262, IEC 62304, DO-178C) requires extensive evidence and documentation, which is difficult for AI/ML [68]. Managing the complexity of interaction with the physical world and ensuring safety across a vast range of operational scenarios is extremely challenging [71]. Exhaustive testing is typically infeasible due to the combinatorial explosion of possibilities [47]. Achieving deterministic replay for debugging and analysis is crucial but difficult [78].
- *Verification Focus:* Emphasis on rigorous methodologies, including formal methods where applicable, extensive simulation-based testing covering edge cases and failure modes, hardware-in-the-loop and real-world testing, fault tolerance analysis, adherence to domain-specific safety standards, meticulous documentation, and end-to-end requirements traceability [65]. Building a robust safety case with sufficient evidence is paramount [78].

Business or mission critical:

- *Requirements:* Business or mission-critical edge AI systems refer to applications that utilise AI to enable real-time decision-making and enhance operational efficiency. These domains demand high levels of scalability, reliability, safety, robustness, and interoperability. Due to their critical nature, they require special attention to ensure performance.
- *Challenges:* Deployment challenges arise from network reliability, as edge devices may operate in environments with unstable connections, affecting data synchronisation and model updates. The diversity of hardware platforms can cause compatibility issues and necessitate tailored solutions. Software challenges include the need for model optimisation, as AI models must be adjusted for edge deployment to balance accuracy and resource utilisation. Environmental conditions also pose risks, as edge devices must withstand various factors that can influence hardware performance and reliability. Regulatory and compliance challenges require navigating global data protection regulations to ensure that AI systems adhere to legal standards.
- *Verification Focus:* Verifying the effectiveness of edge AI systems involves establishing rigorous processes to ensure AI models meet performance standards under diverse conditions. Performance evaluation includes conducting real-time benchmarks to assess the responsiveness, accuracy, and resource utilisation of AI models deployed on edge devices. Interoperability ensures that edge AI solutions operate and communicate effectively within existing ecosystems and various hardware. Compliance verification requires regular audits to ensure that edge AI systems adhere to data privacy laws and industry regulations. Robustness verification involves stress-testing models against adversarial attacks and unexpected inputs to confirm their resilience in real-world scenarios. Lifecycle management strategies are necessary for overseeing the entire lifecycle of edge AI systems, from development and deployment to decommissioning.

Consumer Applications (e.g., E-commerce, social media, entertainment):

- *Requirements:* Often prioritize performance (e.g., accuracy of recommendations, speed of response), user experience, scalability, and cost-effectiveness. While direct physical safety risks are typically lower, significant concerns exist around fairness, bias, privacy, security (e.g., data breaches), misinformation, and ethical use [66].

- *Challenges:* Managing bias and fairness effectively across large, diverse user populations [66]. Protecting user privacy in data-hungry applications. Detecting and mitigating the generation or spread of harmful content or misinformation [67]. Preventing user manipulation (e.g., prompt injection in chatbots) [67]. Understanding and mitigating potential large-scale societal impacts [79][80].
- *Verification Focus:* Often relies more heavily on empirical testing, such as testing for performance, user studies for usability and acceptance, large-scale fairness and bias audits, privacy impact assessments and compliance checks, evaluation of content safety filters, and robustness testing against common failure modes or attacks. While formal verification might be used for specific critical components (e.g., payment processing), the overall verification rigor may be less intense than in safety-critical domains, unless specific high-risk functions are involved.

The fundamental difference in verification approaches between these contexts stems from the level of acceptable risk and the potential severity of failure consequences. Safety-critical domains operate with extremely low risk tolerance, demanding the highest levels of assurance and necessitating the use of more rigorous, often formal, verification techniques, alongside adherence to strict regulatory frameworks [65]. Consumer applications, while facing significant ethical and societal risks, typically have a higher tolerance for certain types of failures (e.g., a poor recommendation vs. a medical misdiagnosis), allowing for a greater reliance on empirical testing and monitoring. Addressing these domain-specific challenges in business and mission-critical edge AI systems is key for ensuring their reliability and effectiveness.

The inherent difficulties in verifying both pure DL (opacity) and pure symbolic AI (brittleness) have spurred interest in hybrid neuro-symbolic approaches [64]. By integrating the pattern-recognition strengths of neural networks with the explicit reasoning and transparency of symbolic methods, these approaches offer a potential pathway to building edge AI systems that are more amenable to verification and trust, particularly for complex tasks [77]. The verification of hybrid systems introduces its own set of research questions regarding the interaction and consistency between the different components [64].

For validation a one-size-fits-all approach to edge AI is ineffective due to the diversity of edge AI technologies and their application domains. The specific validation focus, methods, metrics, and acceptance criteria must

be tailored to the type of AI system and the context in which it operates, particularly considering the nature of user interaction and potential real-world consequences.

Validation Nuances Across AI Types

Generative AI (e.g., LLMs, SLMs, VLMs, image generators): Validation priorities include assessing factual accuracy (mitigating "hallucinations"), ensuring content safety (detecting toxicity, bias, harmful content), preventing malicious use (e.g., generating disinformation or deepfakes), and evaluating output quality attributes like coherence, relevance, and creativity, which often lack objective metrics [66]. The inherent non-determinism is a key challenge, requiring validation strategies that assess output distributions or use human evaluation and red-teaming [81]. Defining and validating the "intended purpose" for highly flexible generative models is complex [87].

Agentic AI and AI agents: The AI agent act as a deterministic component with limited scope, while agentic AI reflects distributed intelligence, characterised by goal decomposition, inter-agent communication, and contextual adaptation, demonstrating key characteristics of the modern agentic AI frameworks. Agentic AI systems define an emergent class of intelligent architectures in which multiple specialised agents collaborate to achieve complex, high-level objectives utilising collaborative reasoning and multi-step planning [17]. V&V of edge AI systems employing AI agents focuses on the reliability and safety of the agent's actions in its operational environment to ensure the agent's decision-making logic is robust and predictable under a broad range of inputs, especially unexpected or anomalous sensor data. This involves rigorous testing of the agent's software, hardware, edge AI algorithms and data components to confirm they meet design specifications and performance benchmarks. Another aspect of V&V for edge AI systems that must be considered is the formal verification of the agent's reasoning processes, which involves creating mathematical models of the agent and its environment to demonstrate that specific critical properties, such as safety, robustness, and resilience, are met. For complex, learning-based agents, this can be supplemented with extensive simulation-based testing to explore the vast state space and identify potential failure modes before deployment in the domain applications.

Autonomous Systems (e.g., autonomous vehicles, industrial robots): Validation overwhelmingly focuses on safety, reliability, and robustness

within complex and dynamic physical environments [7]. Key challenges include achieving sufficient test coverage across a vast space of potential scenarios (combinatorial explosion), bridging the gap between simulation and real-world performance, validating perception systems, and ensuring safe decision-making under uncertainty [78]. Validation heavily relies on extensive simulation, structured scenario-based testing, field testing, formal methods for safety-critical properties, and potentially runtime verification/monitoring [78]. Validating human oversight mechanisms is also critical, especially in military or safety-critical contexts [71].

Decision Support Systems (e.g., medical diagnosis aids, credit scoring tools): Validation emphasizes accuracy, reliability, fairness, explainability, and the system's impact on human decision-making and outcomes [7]. Challenges include validating against potentially imperfect or subjective ground truth, ensuring edge AI recommendations are beneficial and not misleading, rigorously assessing and mitigating bias across different user groups, and providing sufficient transparency to enable user trust and accountability. Validation typically requires domain-specific performance metrics, evaluation by domain experts, user studies assessing impact on decisions, and thorough bias and fairness audits [85].

Domain-Specific Considerations

The application domain significantly shapes validation priorities and methods due to differing risk profiles, regulatory requirements, and stakeholder concerns:

Healthcare: Extremely high stakes due to direct impact on patient safety and well-being [82]. Validation must adhere to regulatory frameworks (e.g., FDA regulations for medical devices, HIPAA for privacy, EU AI Act classifying medical AI as high-risk). Key validation elements include clinical efficacy (proven through clinical evaluation/trials), safety, reliability, data privacy, mitigation of bias in diverse patient populations, usability for clinicians, and explainability to support clinical judgment and trust. Frameworks like FUTURE-AI offer specific guidance for trustworthy AI in healthcare [86].

Finance: Focus on regulatory compliance (e.g., financial conduct authorities, anti-discrimination laws), fairness and bias mitigation in areas like credit scoring and loan applications, accuracy in fraud detection, model risk management, robustness against market volatility, security against financial attacks, and explainability for audits and customer inquiries [83]. Validation involves rigorous back testing, stress testing under various market conditions,

comprehensive bias audits using relevant fairness metrics, security penetration testing, and checks for regulatory adherence.

Transportation (especially Autonomous Vehicles): Safety is the absolute priority [82]. Validation must demonstrate safe operation under a vast range of environmental conditions (weather, lighting, road types) and interactions (other vehicles, pedestrians, cyclists). This involves validating perception systems (sensor fusion, object detection/classification), prediction models, and planning/control algorithms [71]. Validation relies heavily on extensive simulation covering millions of virtual miles, structured scenario-based testing (including edge cases and failure modes), real-world road testing, and the development of robust safety cases supported by evidence [78]. Formal verification methods may be applied to critical safety properties [71].

Social media / Content Platforms: Key concerns involve mitigating the spread of misinformation and harmful content, addressing algorithmic bias in content ranking and recommendation, ensuring fairness in content moderation, protecting user privacy, and managing the impact on user well-being and societal discourse [84]. Validation is challenging due to the massive scale, the dynamic nature of content and user behaviour, the subjectivity involved in defining "harmful" or "fair," and the difficulty in measuring long-term societal impacts. Validation methods often include large-scale testing, human content review and rating, analysis of user engagement and feedback data, and monitoring metrics related to bias, toxicity, and content diversity.

This context-dependency highlights that effective AI validation requires not only technical expertise, but also deep domain knowledge [86]. Generic validation checklists are insufficient; protocols must be tailored to the specific AI type, its intended application, the operational environment, the relevant risks, and the specific needs and values of the stakeholders in that domain [88].

1.8 Trends and Methodologies in Edge AI Verification and Validation

The field of edge AI verification is rapidly evolving, driven by the increasing capabilities and deployment of AI systems, alongside growing concerns about their trustworthiness and potential risks. The unique challenges of edge AI validation are driving significant research and development into new methodologies, techniques, and tools. These efforts aim to provide more rigorous,

scalable, and comprehensive ways to ensure edge AI systems are fit for their intended purpose.

Formal methods are advancing with a growing interest in applying, mathematically rigorous techniques, to the verification and validation of edge AI systems. Techniques like model checking, theorem proving, abstract interpretation, and reachability analysis are being adapted to prove specific properties of edge AI components, especially neural networks, concerning safety, robustness against perturbations (e.g., adversarial examples), and fairness. Major challenges remain in scaling these methods to handle the high dimensionality and complexity of edge AI models and in formally specifying properties for systems operating under uncertainty or with incomplete requirements. Research focuses on developing more scalable algorithms, better abstraction techniques, and methods for probabilistic verification.

Explainable and interpretable AI for V&V are techniques that are increasingly explored as tools for validation and verification. By providing insights into why a model makes a certain prediction (e.g., identifying important input features using SHAP or LIME, visualizing attention mechanisms, generating counterfactual explanations), AI explainability and interpretability can help validators assess whether the model's reasoning aligns with domain knowledge, requirements, or certain rules or principles (e.g., ethical). The techniques can aid in debugging unexpected behaviours, identifying reliance on spurious correlations, and verifying compliance with constraints (e.g., fairness). This helps address the "black box" challenge for validation purposes. The reliability and interpretation of explanations themselves require validation, and research is ongoing to understand the effectiveness and limitations of using AI explainability and interpretability for V&V tasks. In the context of edge machine vision, lightweight explainability methods can help assess whether the model's attention aligns with relevant image features. These methods assist in verifying that edge vision models respond to semantically appropriate cues and not to background artifacts or compression noise.

Neuro-Symbolic AI combines the strengths of data-driven neural networks (sub-symbolic AI) with rule-based logical reasoning (symbolic AI). The symbolic component can represent explicit domain knowledge, constraints, or reasoning rules, potentially making the hybrid system more interpretable, data-efficient, and robust. From a validation perspective, neuro-symbolic approaches offer promise by potentially enabling formal verification of the symbolic reasoning part, using symbolic knowledge to constrain or validate the neural network's outputs, and providing more transparent explanations

for system behaviour. Research is actively exploring different integration architectures and their implications for validation.

Agentic AI and AI agents brings new challenges required the advancements of research focusing on developing new V&V techniques tailored to the dynamic nature of agentic AI, including advancing methods in runtime monitoring and formal verification that can cope with learning-based components and non-determinism. Creating simulation platforms that can model complex, real-world physics and multi-agent interactions will be crucial for testing edge systems exhaustively before deployment. The use of digital twin and immersive triplet environments could enable the safe exploration of an agent's behaviour under a wide range of standard and adverse conditions, helping to identify potential failure modes early. In this context, based on the technology trends research should address the system-level and collaborative aspects of agentic AI at the edge by creating frameworks for validating not only individual agents but also the collective, emergent behaviour of multi-agent systems. Developing techniques to ensure that the goals of individual agents remain aligned with the overall system objectives, even as they adapt and learn, is paramount. Research into explainable XAI and IAI for edge devices is required, as it will enable human operators to understand, trust, and effectively manage the decisions of autonomous agents, ensuring safe and predictable operation in complex, real-world scenarios.

1.9 Conclusion

The rapid advancement and deployment of edge AI necessitate a parallel evolution in the designers' ability to ensure that edge AI systems are safe, reliable, fair, and aligned with human values. Verification, as defined by standards such as ISO/IEC 22989, is the assurance through objective evidence that specified requirements have been fulfilled, forming a cornerstone of building essential trust. It provides the rigorous checks needed to confirm that AI systems are built according to their intended design and specifications.

The unique characteristics of AI, particularly its potential opacity, non-determinism, complex data dependencies, and difficulty in formally specifying requirements for emergent behaviours, pose significant challenges to traditional V&V approaches. The black-box nature of many models hinders direct inspection, scalability limitations restrict the application of formal methods, and the dynamic nature of edge AI systems and their environments demands continuous evaluation beyond design-time checks. Addressing conceptual challenges related to fairness, value alignment, and

adversarial robustness requires ongoing fundamental research, and significant progress is being made. International standards provide common terminology (ISO/IEC 22989), frameworks for trustworthiness (ISO/IEC TR 24028), and management systems for responsible AI governance (ISO/IEC 42001). Methodologically, an increasing number of researchers are adopting formal methods for specific AI and edge AI verification tasks, developing advanced testing techniques (e.g., metamorphic and adversarial testing).

Metamorphic testing techniques are used to verify the behaviour of AI models, when predicting the exact output for a given input is challenging or impossible. The metamorphic testing techniques focus on identifying relationships between inputs and outputs, known as metamorphic relations, that act as logical rules or properties that should hold true when inputs are modified. Adversarial testing is a technique in which inputs are intentionally designed to expose weaknesses or flaws in a system, thereby identifying scenarios where the system produces harmful or biased outputs. This enables testers to identify vulnerabilities and ensure the system responds safely and effectively [37].

Generative AI excels at pattern recognition, classification, and predictive analytics, generates new patterns and multimodal content (e.g., text, sound, images) and plays a dual role in the verification and validation process, for example, as part of an edge AI system that has to be verified and validated and as a technology that supports the V&V processes by generating V&V requirements, specifications and automatically performing the V&V.

The V&V of emerging edge AI agents face challenges arising from the inherent autonomy and the dynamic environments in which these agents operate. The agents can rely on machine learning models that can produce non-deterministic outputs, making the behaviour difficult to predict and formally verify. Continuous interaction with external environments introduces an extensive and unpredictable operational space, where unforeseen events can lead to emergent behaviours that were not anticipated during the design and testing phases, posing risks to safety and reliability.

Further challenges for the V&V processes are the unique constraints of the edge environment itself. Edge AI systems must operate within the limitations of computational power, memory, and energy, which can impact the performance and consistency of their decision-making processes. Edge AI agents must make real-time decisions, where latency is a critical factor. Validating that an edge AI agent responds correctly and within strict time constraints, especially when facing intermittent connectivity or degraded

sensor input, is a significant hurdle that requires novel testing methodologies beyond traditional software V&V.

The process of V&V agentic edge AI systems requires addressing the interaction between human and AI agent components to ensure that the agent's behaviour is understandable and transparent to human users, which allows for efficient oversight and human intervention when necessary. The V&V process must confirm that the edge system can communicate its state and intentions evidently and that its autonomous actions are auditable, interpretable and explainable.

This supports explainable AI techniques, implementing runtime verification for operational assurance, and opening the use of neuro-symbolic architectures to bridge the gap between learning and reasoning. Neuro-symbolic AI is a type of AI that integrates neural and symbolic AI architectures to address the weaknesses of each, providing a robust AI capable of reasoning, learning, and cognitive modelling.

Continued research is essential to develop more scalable and robust verification techniques that can handle the complexity of new edge AI systems. Addressing foundational AI safety problems, enhancing automated and human-centric V&V approaches, and building comprehensive, trustworthy AI frameworks that integrate technical verification with ethical considerations and governance are key priorities. Achieving verifiably trustworthy AI requires a holistic perspective, acknowledging the interplay between hardware, software, AI models, data, systems, processes, and the technical, application, and environmental contexts [40].

Achieving the goal of trustworthy AI and edge AI systems that are demonstrably beneficial and responsibly integrated into society requires elevating validation beyond a mere technical, end-of-phase check. It demands a holistic, continuous, and lifecycle-integrated approach [54].

This approach must rigorously integrate technical validation (ensuring robustness, reliability, and security) with user-centric validation (confirming usability, fitness for purpose, meeting needs), ethical validation (assessing fairness, accountability, and value alignment), and real-world effectiveness monitoring [90].

Success requires multidisciplinary collaboration, bringing together AI/ML experts, software engineers, domain specialists, human factors engineers, ethicists, social scientists, legal experts, end-users, and regulators. International standard bodies like ISO and organisations like NIST provide essential frameworks, common terminology, and guidance (e.g., ISO 9000,

ISO/IEC/IEEE 15288, ISO/IEC 22989, ISO/IEC TR 24028, ISO/IEC 42001, NIST AI RMF) [89].

The rapidly evolving nature of AI means that significant ongoing research and innovation in validation techniques are imperative to address the challenges effectively.

Edge AI system validation is a dynamic and increasingly critical field. As AI capabilities continue to advance and these systems become more deeply embedded in our lives, the methods used to ensure they are fit for purpose, safe, and aligned with human values must also evolve.

The focus is shifting from static, pre-deployment checks towards continuous, adaptive, and context-aware validation processes that span the entire edge AI lifecycle. Addressing the complex technical, ethical, and societal challenges associated with edge AI validation requires sustained research, multidisciplinary collaboration, and international cooperation.

Continued innovation in validation methodologies and tools will be essential to harness the transformative potential of AI responsibly and build a future where edge AI systems are trustworthy and integrated into industrial and business processes.

Acknowledgements

This publication has received funding through the projects Chips JU EdgeAI and HE dAIEDGE. The Chips JU EdgeAI "Edge AI Technologies for Optimised Performance Embedded Processing" project is supported by the Chips Joint Undertaking and its members including top-up funding by Austria, Belgium, France, Greece, Italy, Latvia, Netherlands, and Norway under grant agreement No 101097300. The HE dAIEDGE "A network of excellence for distributed, trustworthy, efficient and scalable AI at the Edge" project is supported under grant agreement No 101120726. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Chips Joint Undertaking. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] A. Lavin et al., "Technology readiness levels for machine learning systems," Nature Communications, vol. 13, no. 1, p. 6039, Oct. 2022, https://doi.org/10.1038/s41467-022-33128-9.

[2] S. Mahmud, S. Saisubramanian, and S. Zilberstein, "Verification and Validation of AI Systems Using Explanations," *Proceedings of the AAAI Symposium Series*, vol. 4, no. 1, pp. 76–80, Nov. 2024, https://doi.org/10.1609/aaaiss.v4i1.31774.

[3] "Verification and Validation of Systems in Which AI is a Key Element - SEBoK," *sebokwiki.org*. https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element.

[4] "ISO/IEC TR 24028:2020 – Overview of trustworthiness in artificial intelligence," *BSI*, 2020. https://www.bsigroup.com/en-IN/training-courses/isoiec-tr-240282020--overview-of-trustworthiness-in-artificial-intelligence/.

[5] "ISO/IEC TR 24028:2020 - Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence" *ISO*. https://www.iso.org/standard/77608.html.

[6] NIST, "AI Risk Management Framework," *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, vol. 1, Jan. 2023, https://doi.org/10.6028/nist.ai.100-1.

[7] J. Jeon, "Standardization Trends on Safety and Trustworthiness Technology for Advanced AI", 2024, https://arxiv.org/abs/2410.22151.

[8] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge, "Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence," *arXiv (Cornell University)*, Feb. 2024. Available at: https://doi.org/10.48550/arxiv.2402.09880

[9] "ISO/IEC 22989:2022 – Information technology – Artificial intelligence – Artificial intelligence concepts and terminology," Edition 1, 2022. https://www.iso.org/standard/74296.html

[10] "Recommendation of the Council on Artificial Intelligence," OECD Legal Instruments, 2025. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

[11] "What is catastrophic forgetting?" IBM, April 2025. https://www.ibm.com/think/topics/catastrophic-forgetting

[12] T. Meuser, et.al. "Revisiting Edge AI: Opportunities and Challenges". IEEE Internet Computing, vol. 28, July-August 2024. https://www.computer.org/csdl/magazine/ic/2024/04/10621659/1Z5lGDb639C

[13] Z. Ren and C. J. Anumba, "Multi-agent systems in construction–state of the art and prospects," Automation in Construction, vol. 13, no. 3, pp. 421–434, 2004, https://doi.org/10.1016/j.autcon.2003.12.002.

[14] G. Papagni, J. de Pagter, S. Zafari, M. Filzmoser, and S. T. Koeszegi, "Artificial agents' explainability to support trust: considerations on

timing and context," AI & Society, Vol. 38, No. 2, pp. 947–960, 2023. https://doi.org/10.1007/s00146-022-01462-7

[15] P. Wang and H. Ding, "The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance," Information Processing & Management, Vol. 61, No. 4, p. 103732, 2024. https://doi.org/10.1016/j.ipm.2024.103732.

[16] F. Sado, C. K. Loo, W. S. Liew, M. Kerzel, and S. Wermter, "Explainable Goal-driven Agents and Robots - A Comprehensive Review", ACM Computing Surveys, Volume 55, Issue 10, pp. 1–41, 2023. https://doi.org/10.1145/3564240.

[17] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenge," arXiv.org, 2025. https://arxiv.org/abs/2505.10468.

[18] C. Riedl and D. De Cremer, "AI for collective intelligence," Collective Intelligence, vol. 4, no. 2, Apr. 2025, https://doi.org/10.1177/26339137251328909.

[19] F. Piccialli, D. Chiaro, S. Sarwar, D. Cerciello, P. Qi, and V. Mele, "AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0," Expert Systems with Applications, vol. 291, p. 128404, Oct. 2025, https://doi.org/10.1016/j.eswa.2025.128404.

[20] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, "A-MEM: Agentic Memory for LLM Agents," arXiv.org, 2025. https://arxiv.org/abs/2502.12110.

[21] D. B. Acharya, K. Kuppan and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," in IEEE Access, vol. 13, pp. 18912-18936, 2025, https://www.doi.org/10.1109/ACCESS.2025.3532853.

[22] R. Zhang et al., "Toward Agentic AI: Generative Information Retrieval Inspired Intelligent Communications and Networking," arXiv.org, 2025. https://arxiv.org/abs/2502.16866.

[23] M. Gridach, J. Nanavati, K. Zine, L. Mendes, and C. Mack, "Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions," arXiv.org, 2025. https://arxiv.org/abs/2503.08979.

[24] E. Miehling et al., "Agentic AI Needs a Systems Theory," arXiv.org, 2025. https://arxiv.org/abs/2503.00237.

[25] S. Hong et al., "MetaGPT: Meta Programming for Multi-Agent Collaborative Framework," arXiv.org, Aug. 07, 2023. https://arxiv.org/abs/2308.00352.

[26] U. M. Borghoff, P. Bottoni, and R. Pareschi, "Human-artificial interaction in the age of agentic AI: a system-theoretical approach," Frontiers in Human Dynamics, vol. 7, May 2025, https://doi.org/10.3389/fhumd.2025.1579166.

[27] J. Heer, "Agency plus automation: Designing artificial intelligence into interactive systems," Proceedings of the National Academy of Sciences, Vol. 116, No. 6, pp. 1844–1850, 2019. https://doi.org/10.1073/pnas.1807184115.

[28] E. Oliveira, K. Fischer, and O. Stepankova, "Multi-agent systems: which research for which applications," Robotics and Autonomous Systems, vol. 27, no. 1-2, pp. 91–106, 1999, https://doi.org/10.1016/S0921-8890(98)00085-2.

[29] Validation - Glossary | CSRC - NIST Computer Security Resource Center, https://csrc.nist.gov/glossary/term/validation

[30] Verification and validation - Wikipedia, https://en.wikipedia.org/wiki/Verification_and_validation

[31] Design Review, Verification and Validation - Quality Gurus, https://www.qualitygurus.com/design-review-verification-and-validation/

[32] Verification Versus Validation – What's the Difference? - Climedo, https://climedo.de/en/blog/verification-versus-validation-whats-the-difference/

[33] System Validation - SEBoK, https://sebokwiki.org/wiki/System_Validation

[34] Implementing ISO 15288 V&V Processes using the V&V Studio - The Reuse Company, https://www.reusecompany.com/wp-content/uploads/2021/02/VV-Studio-Webinar-Jan-2021.pdf

[35] Verification (glossary) - SEBoK, https://sebokwiki.org/wiki/Verification_(glossary)

[36] "Sapien's AI Glossary of Data Terms, Definitions & Insights," Sapien.io, 2025. https://www.sapien.io/glossary/all

[37] A. Pande, "Metamorphic and adversarial strategies for testing AI systems," Ministry of Testing, Jan 14, 2025. https://www.ministryoftesting.com/articles/metamorphic-and-adversarial-strategies-for-testing-ai-systems

[38] ISO and IEC Make Foundational Standard on Artificial Intelligence Publicly Available, https://www.holisticai.com/news/iso-iec-22989-foundational-standard-on-ai-open-source

[39] The foundational standards for AI | JTC 1, https://jtc1info.org/wp-content/uploads/2022/06/03_08_Paul_Milan_Wei_The-foundational-standards-for-AI-20220525-ww-mp.pdf

[40] E. Manziuk, O. Barmak, I. Krak, O. Mazurets, and T. Skrypnyk, "Formal Model of Trustworthy Artificial Intelligence Based on Standardization," CEUR-WS.org, https://ceur-ws.org/Vol-2853/short18.pdf

[41] ISO/IEC TR 24028:2020 - Information technology - Artificial intelligence - OECD.AI, https://oecd.ai/en/catalogue/tools/isoiec-tr-2402820 20-information-technology-artificial-intelligence-overview-of-trustwo rthiness-in-artificial-intelligence

[42] Exploring the landscape of trustworthy artificial intelligence: Status and challenges, https://content.iospress.com/articles/intelligent-decision-t echnologies/idt240366

[43] ISO 42001 Artificial Intelligence Management System - Amazon Web Services (AWS), https://aws.amazon.com/compliance/iso-42001-faqs/

[44] ISO 42001 - AI Management System - BSI, https://www.bsigroup.com /en-US/products-and-services/standards/iso-42001-ai-management-s ystem/

[45] ISO/IEC 22989:2023 Understanding AI Concepts and Definitions Training Course | BSI, https://www.bsigroup.com/en-ID/training-courses/isoiec-229892023-understanding-ai-concepts-and-definitions -training-course/

[46] AI Compliance Audit: Step-by-Step Guide - Dialzara, https://dialzara.c om/blog/ai-compliance-audit-step-by-step-guide/

[47] R. Prieto, "Verification and Validation of Project Management Artificial Intelligence Key Points," Jun. 2020. https://www.researchgate.net/publi cation/342452507_Verification_and_Validation_of_Project_Managem ent_Artificial_Intelligence_Key_Points

[48] AI audit checklist (updated 2025) | Complete AI audit procedures | Technical evaluation framework | System reliability guide | Compliance checklist | Lumenalta, https://lumenalta.com/insights/ai-audit-checklis t-updated-2025

[49] Y. Wang and S. H. Chung. "Artificial intelligence in safety-critical systems: a systematic review," Industrial Management & Data Systems,| Emerald Insight, Dec. 2021. https://www.emerald.com/insight/content/ doi/10.1108/imds-07-2021-0419/full/html

[50] Trustworthy AI - AI@UCSF - University of California San Francisco, https://ai.ucsf.edu/trustworthy

[51] AI Risks and Trustworthiness - NIST AIRC - National Institute of Standards and Technology, https://airc.nist.gov/airmf-resources/ai rmf/3-sec-characteristics/

[52] S. A. Seshia, D. Sadigh, and S. Shankar Sastry. Toward Verified Artificial Intelligence. Communications of the ACM, July 2022. https://cacm.acm.org/research/toward-verified-artificial-intelligence/
[53] AI, Opacity, and Personal Autonomy, https://d-nb.info/1275205275/34
[54] Measure - NIST AIRC - National Institute of Standards and Technology, https://airc.nist.gov/airmf-resources/playbook/measure/
[55] Understanding the NIST AI RMF: What It Is and How to Put It Into Practice - Secureframe, https://secureframe.com/blog/nist-ai-rmfy
[56] AI Life Cycle Core Principles - CodeX - Stanford Law School, https://law.stanford.edu/2023/03/17/ai-life-cycle-core-principles/
[57] Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research - Nuffield Foundation, https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf
[58] Messages on "When using AI systems, what are some best practices for ensuring the results you receive are accurate, relevant, and aligned with your original goals?" - ProjectManagement.com, https://www.projectmanagement.com/discussion-topic/203772/when-using-ai-systems--what-are-some-best-practices-for-ensuring-the-results-you-receive-are-accurate--relevant--and-aligned-with-your-original-goals-?sort=asc&pageNum=39
[59] A Framework for the Verification and Validation of Artificial Intelligence Machine Learning Systems - JagWorks@USA - University of South Alabama, https://jagworks.southalabama.edu/theses_diss/137/y
[60] Trustworthy AI - The Data Science Institute at Columbia University, https://datascience.columbia.edu/news/2020/trustworthy-ai/
[61] NIST launches ARIA program to assess societal impacts, ensure trustworthy AI systems, https://industrialcyber.co/ai/nist-launches-aria-program-to-assess-societal-impacts-ensure-trustworthy-ai-systems/
[62] User Acceptance Testing (UAT): Definition, Process, and Tools - LambdaTest, https://www.lambdatest.com/learning-hub/user-acceptance-testing
[63] Human-AI Interaction and User Satisfaction: Empirical Evidence from Online Reviews of AI Products - ResearchGate, https://www.researchgate.net/publication/390142284_Human-AI_Interaction_and_User_Satisfaction_Empirical_Evidence_from_Online_Reviews_of_AI_Products/download

[64] J. Renkhoff, K. Feng, M. Meier-Doernberg, A. Velasquez, and H. H. Song, "A Survey on Verification and Validation, Testing and Evaluations of Neurosymbolic Artificial Intelligence," IEEE transactions on artificial intelligence, pp. 1–15, Jan. 2024, https://doi.org/10.1109/tai.2024.3351798.

[65] A. E. Goodloe, "Assuring Safety-Critical Machine Learning-Enabled Systems: Challenges and Promise," *Computer*, vol. 56, no. 9, pp. 83–88, Sep. 2023, https://doi.org/10.1109/mc.2023.3266860

[66] A. Woodie, "Top 10 Challenges to GenAI Success," BigDATAwire, Jan. 22, 2024. https://www.bigdatawire.com/2024/01/22/top-10-challenges-to-genai-success/

[67] C. Bronsdon, "AI Safety Metrics: How to Ensure Secure and Reliable AI Applications" - Galileo AI, 2025, https://www.galileo.ai/blog/introduction-to-ai-safety

[68] R. Camacho, "A Practical Guide for AI in Safety-Critical Embedded Systems - Parasoft, 2025, https://www.parasoft.com/blog/ai-in-safety-critical-embedded-systems/

[69] Y. Y. Elboher et., al "Formal Verification of Deep Neural Networks for Object Detection," Arxiv.org, 2023. https://arxiv.org/html/2407.01295

[70] What are non-deterministic AI outputs? - Statsig, 2024, https://www.statsig.com/perspectives/what-are-non-deterministic-ai-outputs-

[71] K. Leahy et al., "Grand Challenges in the Verification of Autonomous Systems," arXiv (Cornell University), Nov. 2024, https://doi.org/10.48550/arxiv.2411.14155.

[72] Symbolic AI vs. Deep Learning: Key Differences and Their Roles in AI Development, https://smythos.com/ai-agents/agent-architectures/symbolic-ai-vs-deep-learning/

[73] Symbolic AI vs. Machine Learning: A Comprehensive Guide - SmythOS, https://smythos.com/ai-agents/ai-tutorials/symbolic-ai-vs-machine-learning/

[74] O. Vermesan, V. Piuri, F. Scotti, A. Genovese, R. D. Labati, and P. Coscia, "Explainability and Interpretability Concepts for Edge AI Systems," River Publishers eBooks, pp. 197–227, Feb. 2024, https://doi.org/10.1201/9781003478713-9.

[75] What Is Explainable AI (XAI)? Palo Alto Networks, https://www.paloaltonetworks.com/cyberpedia/explainable-ai

[76] Deep Learning's Challenges and Neurosymbolic AI's Solutions - AskUI, 2024. https://www.askui.com/blog-posts/deep-learnings-challenges-and-neurosymbolic-ais-solutions

[77] V. Musanga, S. Viriri, and C. Chibaya, "A Framework for Integrating Deep Learning and Symbolic AI Towards an Explainable Hybrid Model for the Detection of COVID-19 Using Computerized Tomography Scans," Information, vol. 16, no. 3, p. 208, Mar. 2025, https://doi.org/10.3390/info16030208.

[78] X. Zhang, "Apex.OS: Breaking Barriers in Autonomous Verification & Validation", 2024, https://www.apex.ai/post/apex-os-breaking-barriers-in-autonomous-verification-validation

[79] J. Szarmach, NIST: Reducing Risks Posed by Synthetic Content An Overview of Technical Approaches to Digital Content Transparency, 2025, https://www.aigl.blog/nist-reducing-risks-posed-by-synthetic-content-an-overview-of-technical-approaches-to-digital-content-transparency/

[80] NIST Trustworthy and Responsible AI - NIST AI 100-4. Reducing Risks Posed by Synthetic Content. An Overview of Technical Approaches to Digital Content Transparency. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf

[81] NIST. The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals, 2024, https://www.nist.gov/document/aisi-strategic-vision-document

[82] S. Dahaweer, "How AI is going to revolutionize safety in critical applications", Alithya, 2023, https://www.alithya.com/en/insights/blog-post/how-ai-going-revolutionize-safety-critical-applications

[83] The Top 10 Unsolved Challenges in AI: A 2024 Retrospective, gekko, 2024, https://gpt.gekko.de/unsolved-challenges-in-ai-2024/

[84] Risks from AI - An Overview of Catastrophic AI Risks, CAIS - Center for AI Safety, https://www.safe.ai/ai-risk

[85] R. Lubecki, "Verifying and Validating AI/ML" - UpCity, 2021, https://upcity.com/experts/verifying-and-validating-ai-ml/

[86] K. Lekadir et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, p. e081554, Feb. 2025, https://doi.org/10.1136/bmj-2024-081554.

[87] Traditional AI vs. Generative AI: What's the Difference? - College of Education, Illinois, 2024, https://education.illinois.edu/about/news-events/news/article/2024/11/11/what-is-generative-ai-vs-ai

[88] N. Ahsan, "Why Enterprises Are Adopting Domain-Specific AI Agents", Vidizmo, 2025, https://vidizmo.ai/blog/why-domain-specific-ai-agents-are-key-to-business-success

[89] Future of AI Research – AAAI – Association for the Advancement of Artificial Intelligence 2025. Available: https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf

[90] Methods For Verifying AI Reliability to Ensure AI Safety, IAI – Institute for AI Transformation, August 2024. Available: https://www.leadersinaisummit.com/insights/methods-for-verifying-ai-reliability-to-ensure-ai-safety

[91] IEEE 1012-2016. IEEE Standard for System, Software, and Hardware Verification and Validation. https://webstore.ansi.org/standards/ieee/ieee10122016?source=blog

[92] O. Vermesan, K. De Bosschere, T. Vardanega, S. Azaiez, M. Duranton, R. Badia, and D. Lezzi (Eds.). "Distributed Computing and Swarm Intelligence - Developing a Vision for Transatlantic Collaboration," Zenodo, Feb. 2025, https://doi.org/10.5281/zenodo.14940197