Evaluation of Robustness, Reliability, and Safety of an Artificial Intelligence Based System

Lucas Mattioli
IRT SystemX, Onera
France
lucas.mattioli@irt-systemx.fr

Annia Abtout
Talan,
France
annia.abtout@talan.com

Martin Gonzalez, Afef Awadid, Kevin Mantissa, Faouzi Adjed
IRT SystemX,
France
{*first-name.last-name*}*@irt-systemx.fr*

Joseph Machrouh, Jaime De Oliveira
Thales Land & Air Systems,
France
{*first-name.last-name*}*@thalesgroup.com*

Christophe Guettier, Hatem Hajri
Safran,
France
{*first-name.last-name*}*@safrangroup.com*

Juliette Mattioli
Thales SA, cortAIx,
France
juliette.mattioli@thalesgroup.com

***Abstract*—This paper proposes a paradigm-aware framework for evaluating AI robustness, reliability, and safety, arguing that current methods, designed for supervised learning, fail to reflect the diversity of modern AI. It distinguishes but links the three properties: robustness as performance under perturbations and distribution shifts; reliability as consistency and calibration over time; and safety as a broader socio-technical goal that depends on but goes beyond both. The main contribution is a systematic analysis of how these properties affect the AI system life cycle defined by ISO/IEC 5338, and how they vary across four paradigms: data-driven, symbolic, hybrid, and generative AI. Each paradigm has characteristic failure modes requiring tailored assessments—from adversarial testing and drift detection for neural networks, to formal verification for symbolic systems, to red-teaming and alignment for large language models. The framework embeds these assessments into systems engineering life cycles, stressing that reliability must be addressed at every stage, from requirements to post-deployment monitoring.**

Keywords- trustworthy AI; robustness; reliability; safety.

I. Introduction

Artificial Intelligence (AI) from major research labs and tech companies is now critical in high-stakes sectors such as aerospace, healthcare, automotive, and defense, where failures can be severe. This has spurred global efforts to define and ensure "trustworthy AI," as rigorously outlined by the European Commission's High-Level Expert Group [1]. Trustworthy AI covers human oversight, technical safety, privacy, transparency, fairness, societal and environmental responsibility, and accountability. In 2024, the EU AI Act made many of these requirements legally binding for high-risk AI systems, while organizations such as EASA and ISO/IEC created standards and certification frameworks for AI safety and reliability. The French program "Confiance.ai" [2] introduced a structured methodology for assessing machine learning trustworthiness, now maintained by the European Trustworthy AI Association, which offers open-source tools for scalable and secure AI development.

This paper introduces a paradigm-aware framework for evaluating the Robustness, Reliability, and Safety (RRS) of AI systems, contending that current, supervised-learning-centric methods overlook other paradigms (see Table I). It shows in section III how RRS concerns differ for data-driven, symbolic, hybrid, and generative AI, each requiring specific evaluations: adversarial testing and drift detection for data-driven models, formal verification for symbolic systems, compositional checks for hybrid models, and red-teaming and alignment assessments for generative AI such as large language models. Then in section IV, RRS is integrated across the AI lifecycle, from specification and design to validation, deployment, and monitoring, emphasizing continuous management in real-world use. The paper concludes in section VI that RRS are systemic properties of deployed AI, not merely features of trained models, and thus demand ongoing reassessment to preserve long-term trustworthiness.

II. Definitions and Conceptual approach

Building on the EU AI Act, the AI HLEG guidelines [1], the Confiance.ai methodology [3], and the EASA AI certification roadmap [4], we define AI trustworthiness through seven pillars: Robustness, Reliability, Safety, Explainability, Fairness, Privacy, and Governance (Table II). This article addresses only Robustness, Reliability, and Safety (RRS), which here have domain-specific meanings beyond conventional systems and software engineering. They are essential for trustworthy AI in critical domains such as healthcare, automotive, aerospace, defence, and security, and must be precisely specified to preserve intended behaviour across diverse scenarios.

A. Robustness, Reliability, and Safety Properties

In traditional engineering, **robustness** is a system's ability to perform reliably under minor disturbances. In AI, robustness must extend beyond adversarial attacks—crafted inputs that mislead models—to include distributional shift, where real-world data differs from training data and degrades performance. For example, a vision-based neural network trained in well-lit conditions may perform well in the lab but fail in dynamic environments. Such failures show that AI robustness requires

TABLE I. AI PARADIGMS W.R.T DOMINANT FAILURE MODES

Paradigm	Examples	Dominant Failure Modes
Data-Driven AI	Deep neural networks, SVM	Adversarial attacks, distributional shift
Symbolic AI	Expert systems, logic programming	Incomplete rule bases, logical inconsistencies
Hybrid AI	Neuro-symbolic, Physics/Geometry-Informed NN	Interface mismatches between constituents
Generative AI	LLMs, diffusion models	Hallucinations, prompt injection, bias

adaptation to varying contexts, not just resistance to noise, and that models appearing robust in tests may be fragile in practice, calling for evaluations beyond traditional ones.

We measure AI **reliability** using metrics such as failure rate, event rate, and error rate. An AI system is reliable if it consistently performs its intended function over time and under specified conditions. Assessment must include hardware failures, where physical components stop working (e.g., a faulty GPU), and software failures, where the system does not fulfil its purpose. Because failures may not halt use, we define 'failure' broadly, allowing multiple failures per unit. Reliability metrics then capture whether the system failed, time to failure, and, for recurring failures, the failure event rate.

Machine learning introduces further dimensions of reliability due to its probabilistic and opaque nature, including failure rates, output correctness, confidence calibration, and temporal stability [5]. A model may score well on a static test set yet still cause critical errors if its confidence is mis-calibrated. Reliability also degrades as data, concepts, or user behaviour change; for example, a predictive maintenance model may begin accurate but drift as its training data become unrepresentative. Reliability therefore demands continuous monitoring and recalibration throughout the system lifecycle.

Under ISO 26262, **safety** goes beyond harm prevention to cover AI-specific risks in dynamic, human-centred settings. In AI systems, safety is a property that reflects model performance, user interaction, misuse, and edge cases. It requires a holistic approach to human-AI collaboration, fail-safe mechanisms, and ethical alignment, supported by technical safeguards, organizational and procedural controls, and human-override protocols for transparent decisions and continuous, adaptive risk assessment and monitoring [6].

Robustness, reliability, and safety are interdependent. At the AI Constituent (AIC) level, robustness, uncertainty, and monitoring cannot generally be cleanly modularized, as they are all built, trained, and calibrated around the AIC's central model. The RUM Methodology [7] provides a principled basis for this view. Unlike model-centric evaluation, it treats AICs as atomic units whose behaviour must be assessed across their full lifecycle—specification, development, deployment, and updating—in line with end-to-end system engineering. It justifies treating AICs as indivisible and provides a structured set of mostly non-aggregative trust metrics to capture trustworthiness across the lifecycle. The framework also offers operational tools, such as AI Blueprints for runtime monitoring, human-in-the-loop interaction, and long-term maintainability, supporting trustworthy AI deployment in industrial settings.

B. Inter-dependencies of Robustness, Reliability and Safety

Robustness, reliability, and safety in AI are closely linked, enabling systems to work in both controlled and real-world settings. This RRS triad must be balanced and evaluated together: robustness handles variation and adverse conditions; reliability ensures stable, repeatable performance; safety limits behaviour to acceptable risk levels. Neglecting any one dimension can create vulnerabilities (*e.g.*, a model that seems reliable in testing may fail in deployment). As AI systems grow more complex and autonomous, jointly assessing these three aspects is vital for technically sound, trustworthy, and socially beneficial deployments.

Robustness is a prerequisite for reliability: A system must maintain performance under unexpected changes or disruptions. An AI model that withstands adversarial inputs, sensor noise, or data shifts is more reliable. Without robustness, performance can degrade unpredictably, causing inconsistent outputs and undermining even well-trained models when small input changes trigger cascading errors.

Reliability alone does not guarantee safety. Technical metrics matter little without systemic context. An AI can be highly reliable—consistently correct under defined conditions—yet still be dangerous if its objectives conflict with safety or operational standards. A language model, for example, may reliably generate fluent, relevant text while still producing harmful, biased, or misleading content if misaligned with safety constraints. Reliability is necessary but not sufficient for safety: a system can function flawlessly yet pose unacceptable risks if its goals are flawed or safeguards against misuse and unintended consequences are missing.

Safety is the culmination of these efforts, extending from technical guarantees to the socio-technical contexts in which AI operates. It is not inherent to a model but emerges from its interactions with environments, users, and systems. A system may perform well under normal conditions yet still cause disasters in rare edge cases—for instance, a healthcare diagnostic AI might excel in trials but become unsafe if clinicians over-rely on it without understanding its limits.

Safety demands a holistic approach that integrates technical robustness, human factors, organizational protocols, and ethics so AI aligns with human expectations, societal norms, and real-world complexity. Robustness, reliability, and safety are hierarchical yet interdependent: robustness resists disruptions, reliability sustains consistent performance, and safety embeds both in human-centered design. Neglecting any produces technically capable but untrustworthy systems, so AI must be evaluated on all three to remain safe and dependable in dynamic real-world settings [8].

TABLE II. ROBUSTNESS, RELIABILITY AND SAFETY ARE PILLARS OF AI TRUSTWORTHINESS

Pillar	Core Question	Key Evidence Types	Primary Standards
Robustness	Does the system maintain performance under perturbations, distributional shift, and adversarial inputs?	Adversarial test sets, OOD benchmarks, stress tests	ISO/IEC 24029, EASA DAL robustness requirements
Reliability	Does the system produce correct outputs consistently across its operational domain?	Performance metrics, failure rate analysis, uncertainty quantification	DO-178C, ISO 26262, IEC 61508
Safety	Does the system avoid causing harm to people, assets, or the environment?	Hazard analysis, FMEA/FMEDA, runtime monitoring logs	ARP 4754B, ARP 4761A, ARP 6983, MIL-STD-882

III. PARADIGM-SPECIFIC ASSESSMENT METHODS

Assessing an AI system's trustworthiness requires understanding its core properties, since AI paradigms differ in architecture, outputs, and failure modes. Four main paradigms have distinct features that shape their reliability.

A. Overview of AI Paradigms

Data-driven AI uses statistical methods like neural networks and evolutionary algorithms to learn from data. It performs well in controlled settings but struggles with real-world unpredictability due to sensitivity to data shifts and attacks. Even with explainability and robustness tools, a gap persists between controlled and real-world performance, leading to bias, failures, and vulnerabilities.

Symbolic AI uses human-readable knowledge bases and logical deduction for interpretability and formal correctness, but its reliability degrades with incomplete or inaccurate data. Missing rules, contradictions, and other gaps cause rigid reasoning. Traditional verification helps, yet maintaining complete, consistent knowledge bases is a core challenge.

Hybrid AI combines approaches such as neuro-symbolic models or constraint-informed neural networks to exploit their strengths. However, component interactions can introduce trust issues, as interface errors may compromise reasoning, even though symbolic constraints can also enhance trustworthiness by restricting outputs to logically valid conclusions.

Generative AI (GenAI), including large language models and diffusion-based image generators, uses vast datasets to create content, but its unpredictability makes it unreliable. It can produce factual errors, be manipulated, and its scale and emergent behaviour undermine traditional verification. Red-teaming and behavioural benchmarks offer partial safeguards, but full reliability remains out of reach.

B. Assessment Methods by Paradigm

RRS assessment in AI systems depends on the paradigm, as each has distinct behaviours, failure modes, and assurance needs. One-size-fits-all methods are ineffective; tailored approaches are essential. Without them, critical vulnerabilities may stay hidden, producing systems that work in controlled settings but fail in real-world use.

Data-driven AI learns statistical patterns from data rather than explicit rules, creating robustness, reliability, and safety challenges. Small input changes can trigger large output shifts (adversarial vulnerabilities), revealed through adversarial testing that simulates worst cases. Robustness certification checks that outputs remain stable within defined ranges. Distributional shift—when deployment data differs from training data—demands continuous monitoring to prevent degradation. Reliability is typically assessed with methods like cross-validation, while safety uses adapted frameworks such as System-Theoretic Process Analysis (STPA). However, most methods are statistical, not causal, limiting failure explanation and systemic risk prevention.

Symbolic AI uses explicit rules and logic for formal verification and deterministic reasoning, supporting robustness by avoiding adversarial vulnerabilities. It faces issues like logical inconsistencies and incomplete knowledge bases that undermine reliability. Formal methods (model checking, theorem proving) give strong correctness guarantees. Safety can benefit from traceable hazard paths, but depends on knowledge-base completeness. Symbolic AI offers strong formal guarantees, yet its effectiveness hinges on rule coverage and quality.

Hybrid AI combines data-driven and symbolic methods, exploiting their strengths while adding interface risks. Robustness requires validating how symbolic components interpret data-driven outputs. Reliability is evaluated via statistical testing plus formal verification, and safety via contract-based design. Embedding physics-informed or domain-specific constraints further improves robustness.

GenAI is challenging because of opaque, emergent behaviour and open-ended outputs. Traditional robustness tests are inadequate, as failures often reflect misalignment rather than adversarial noise. Prompt robustness testing examines how input variations affect outputs, exposing vulnerabilities like jailbreaking and prompt injection. Watermarking embeds detectable patterns in content to flag AI-generated misinformation [9]. Reliability covers accuracy, truthfulness, bias mitigation, and factuality, supported by metrics and bias audits. The main safety problem is alignment, keeping models beneficial and controllable. Techniques such as red-teaming and Reinforcement Learning from Human Feedback target value alignment but lack formal guarantees, leaving residual safety risks and highlighting the need for new assurance methods.

IV. INTEGRATION INTO SYSTEMS ENGINEERING LIFE-CYCLES

Trustworthiness cannot be added to an AI system after development; it must be built in from the outset as a system-level property spanning the entire lifecycle (see Figure 1 and Table III), from inception through operation and monitoring. This section explains how our approach aligns with established systems engineering practices, focusing on the EASA AI concept paper and ARP 6983 for aerospace [10], the ISO/IEC

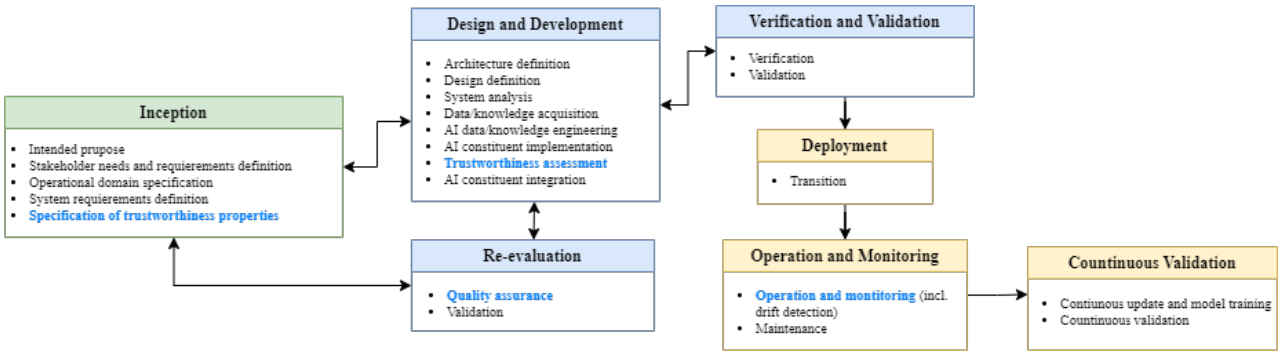

Figure 1. Variation of the ISO/IEC 5338 AI system life cycle processes with respect to robustness, reliability and safety requirements

TABLE III. THE RRS ACTIVITIES W.R.T. THE ENGINEERING PHASES

Phase	RRS Activities
Requirements	Define RRS goals (*e.g.*, "System must handle 95% of adversarial inputs without failure").
Design	Architect for assurance (*e.g.*, modularity for hybrid AI).
Verification	Apply paradigm-specific methods (*e.g.*, formal proof for symbolic AI).
Validation	Test in operational environments (*e.g.*, shadow mode for LLMs). Post-Deployment Continuous monitoring (*e.g.*, drift detection, red-teaming updates).

5338 standard for AI-specific software lifecycle processes[1], and the Confiance.ai methodology for industrial AI [11].

A. Inception Phase: Specifying Trustworthiness Properties

The Inception Phase captures high-level intents, refines them into detailed requirements, and decomposes tasks into executable units with planned components. It underpins all later assessment activities. The AI system must function reliably as intended and remain robust in unexpected situations. Robustness, reliability, and safety requirements must be expressed in measurable, testable terms; vague statements like "the system shall be robust" are not verifiable.

An AI system's **intended purpose** defines its function, outputs, users and performance limits. Set during Inception, it shapes verification and validation by specifying the system's role and acceptable behaviour. The **operational domain** (OD) defines the conditions under which an AI system is designed to function as intended [10]. The OD is part of the functional specification for trustworthiness because: 1) transparency about the OD clarifies system capabilities and limits (as required by the AI Act); 2) the OD is the reference domain for all operational trustworthiness attributes; and 3) the OD itself must be complete, consistent and human-readable. The OD is thus both a design constraint and an assessment tool: tests must verify performance throughout the OD, and inputs outside it must trigger defined safe behaviours.

- **Performance requirements**: quantitative reliability targets for each scenario class within the OD and operational needs, validated against the hazard analysis rather than generic engineering judgment.
- **Robustness requirements**: quantitative limits on acceptable performance degradation under defined perturbation types and magnitudes, including adversarial and distributional robustness.
- **Safety requirements**: constraints on AI outputs derived from system-level hazard analysis, expressed as invariants that must never be violated. For data-driven components these are probabilistic; for symbolic components they can be formal logical constraints verifiable by model checking.

The OD defines the system's environment, inputs, and operational constraints, reflecting the data used for training and validation. Intended purpose concerns outputs and overall function, while OD concerns inputs and conditions of use. Deviating from the intended purpose is misuse; operating outside the OD, even for an authorized purpose, creates out-of-distribution scenarios that can degrade performance without warning. The OD sets the system's operational boundaries; the intended purpose sets its accountability and functional limits.

B. Design Phase and Development: RRS Assessment

The design phase defines the system architecture to assess feasibility and later evaluation costs. Sound design principles can improve RRS and simplify its assessment.

- **Separation of concerns**: architectures that separate perception (data-driven), reasoning (symbolic), and actuation allow component-level assessment with paradigm-specific methods and clear interfaces, which is far more tractable than assessing a monolithic end-to-end neural system.
- **Explicit uncertainty representation**: propagating and exposing uncertainty estimates at all interfaces (rather than using point predictions) enables runtime monitoring and lets the symbolic reasoning component act conservatively when the data-driven component is uncertain.
- **Redundancy and diversity**: for high-assurance applications, architectural redundancy with diverse AI implementations can achieve required reliability and safety even when no single AI component can. Dissimilar redundancy—using different datasets, architectures, or paradigms—mitigates common-cause failures that defeat homogeneous redundancy.

[1]https://standards.globalspec.com/std/14651195/iso-iec-5338

- **Graceful degradation**: defining degraded operational modes (fall-forward to simpler, more reliable AI components; fall-back to human operators; fail-safe to a defined safe state) creates an architectural safety barrier that limits the consequences of AI component failure, regardless of failure mode.

C. Verification and Validation Phase

The AI validation and verification (V&V) process follows these principles:

- **Test plan coverage**: the test plan must explicitly cover the OD, sampling all identified scenario classes and oversampling safety-relevant minority classes. Any OD coverage gaps are safety-critical. Test set size must satisfy statistical power requirements for each metric, with documented sample size calculations.
- **Independence of V&V**: the V&V team must be independent from the development team to avoid optimistic bias, as required in safety-critical software standards. The V&V team must not have contributed to training data collection, model development, or hyperparameter selection.
- **Learning process verification**: for data-driven AI and GenAI, V&V must assess both the trained model and the training process. This includes verifying that the dataset meets documented quality requirements, the training pipeline is reproducible, hyperparameter selection did not contaminate the test set, and the trained model matches the documented architecture and configuration.
- **Formal analysis where feasible**: when architecture and computational budget allow, formal verification should supplement statistical testing. For symbolic components, it provides completeness guarantees; for hybrid components, compositional verification combines formal results for symbolic parts with statistical results for neural parts to obtain system-level guarantees.

D. Re-evaluation phase and post deployment monitoring

Pre-deployment V&V offers only point-in-time quality assurance, confirming the system meets requirements at deployment. For AI systems,this is insufficient as performance can degrade with changing operational environments. Continuous assurance, ongoing collection and evaluation of RRS evidence across the lifecycle, is therefore essential.

The **post-deployment monitoring** architecture must be defined in the system design. Key components include: drift monitors comparing operational to training data; monitors of accuracy and failure rates against operational ground truth; detectors for unusual outputs; and pipelines logging safety-relevant events. Monitoring is challenged by AI opacity and emergent capabilities, so improved anomaly detection and model evaluation are needed.

Performance degradation thresholds (quantitative performance drops that trigger alerts, degraded operation, or system withdrawal) must be specified in operational requirements and tied to the safety case. Any AI changes—retraining, dataset updates, configuration changes—must undergo change impact assessment to determine whether partial or full re-verification is required.

V. Examples

The **aeronautics industry**'s growing reliance on AI-based systems—from in-service support and autonomous flight operations to pilot decision support—demands unprecedented rigor in assessing robustness, reliability, and safety. Unlike traditional, deterministic or rule-based aviation software, data-driven or hybrid AI models learn from data, introducing variability that requires specialized validation. For example, machine learning algorithms for computer vision in aircraft inspection, such as Airbus tools that detect micro-cracks in composite materials, must be robust to adversarial inputs (manipulated images or sensor noise) that could cause false negatives and missed structural failures. Predictive analytics tools, like those Safran uses to anticipate engine component wear, must remain accurate with incomplete or biased historical data to support reliable maintenance decisions. Safety-critical avionics, including single-pilot operations and AI-driven flight optimizers (such as the TopSKy Sequencer by Thales [12]), need fail-safe mechanisms for edge cases like severe weather or conflicting air traffic control instructions to preserve passenger safety. Thus, even though traditional aviation standards (*e.g.* DO-178C and ARP 6983) provide partial guidance and EASA is developing a unified AI certification framework for aeronautics [10], dedicated RRS assessments are essential. Deep learning in pilot assistance tools, such as Airbus's Vision Based Landing Approach Runway Detection (LARD) [3], require continuous monitoring to detect 'black box' decision drifts. Without robust verification and validation—such as formal methods or stress testing on synthetic data—even well-trained models may fail in rare but plausible situations, including cyber-attacks.

In the **healthcare domain**, the growing use of AI systems, from data-driven diagnostic tools to generative models for clinical decision support, demands rigorous assessment of RRS. Unlike traditional clinical software based on deterministic rules, data-driven imaging models learn statistical patterns from training data, introducing variability that requires specialized validation. Deep learning algorithms for radiology or pathology must generalize beyond their development site; in practice, sensitivity to input noise, scanner differences, and population shifts often reveals brittleness in models whose internal validation had seemed adequate. Reliability is further weakened by confidence miscalibration, temporal drift in populations and imaging protocols, and inconsistent performance across operating conditions, all of which can cause systematic yet hidden errors. Safety, in turn, goes beyond technical performance: a diagnostic AI may excel in trials yet be unsafe if clinicians over-rely on it, misunderstand its limits, or lack human-override procedures for out-of-scope inputs.

For GenAI in clinical decision support, the RRS profile differs markedly from that of data-driven systems. Robustness is undermined by prompt sensitivity, as fabricated details in clinical prompts can cause models to elaborate on embedded errors; prompt-based safeguards mitigate but do not remove this

risk. Reliability must cover not only accuracy but also factual consistency and temporal stability, since reasoning failures are a primary source of errors and cast doubt on LLM trustworthiness across repeated high-stakes interactions. Safety arises from how model outputs integrate into clinical workflows: clinicians may over-trust fluent but wrong answers, and the lack of clear failure signals creates regulatory obstacles for approval as medical devices. This underscores the need for clinically tailored red-teaming, alignment protocols, and human-override mechanisms as structural assurance requirements.

Assessing robustness, reliability and safety is not only a technical task but a prerequisite for regulation and public trust. Yet major challenges remain. One proposed solution is scalable formal verification. While formal methods offer the strongest guarantees, they are not yet systematically applicable to large neural networks, and current certified robustness techniques yield limited guarantees and reduced accuracy. A key research direction is extending formal verification to hybrid AI via compositional methods, abstraction refinement, and architectures designed for verifiability.

Another important direction is specifying OD for high-dimensional, unstructured inputs. For systems using natural language, raw sensor data or visual scenes, precisely defining the operational design domain is difficult. There is an urgent need for formal OD specification languages for such inputs, and for automated tools for OD boundary detection and coverage measurement.

GenAI safety verification remains largely unsolved. Hallucination and prompt injection in large language models lack systematic verification methods. Manual red-teaming is costly, incomplete and non-reproducible. Automatic red-teaming, formal behavioural specifications for LLMs and rigorous GenAI safety evaluation approaches are open problems.

Current reliability and safety assessments mainly use statistics to measure what a system does, not why. Causal models could reveal root causes, predict failure conditions and greatly strengthen safety cases. Integrating causal reasoning with machine learning is an active research area with important implications for assurance.

Our approach defines multiple assessment dimensions, but combining them into a single assurance claim needs formal methods for aggregating heterogeneous evidence. Research on evidence combination, uncertainty propagation in safety cases, and formal assurance case logics is required.

VI. Conclusion and Future Works

This paper presents the first paradigm-aware approach to evaluating robustness, reliability, and safety of AI systems across data-driven, symbolic, hybrid, and generative AI. It offers: precise operational definitions of these properties; a systematic analysis of how they appear and fail in each paradigm; a survey of assessment methods and evidence types for gap analysis and evidence planning; and a regulatory mapping to the EU AI Act, EASA, DO-178C/ARP 4754A, IEC 61508, and the Confiance.ai end-to-end methodology.

Three conclusions follow. First, assessment must be paradigm-aware: methods for symbolic solvers, deep neural networks, and large generative language models differ, and a uniform approach leaves gaps. Second, robustness, reliability, and safety are interdependent but distinct; each requires separate assessment before integration into a safety case. Conflation creates blind spots: a system reliable in testing may fail under distribution shift, and one robust to perturbations may still be hazardous in some contexts. Third, assurance is a lifecycle activity, not a one-off certification: changing environments, performance drift, and new adversarial techniques demand continual evidence collection, evaluation, and updating. As AI becomes more capable and embedded in critical infrastructure, inadequate assurance becomes more costly. The proposed approach is a step toward an AI assurance engineering discipline that evolves with the technology it governs.

Future work aims to complete the approach defined in [3] going beyond data-driven AI by extending AI engineering methods and tools to symbolic, hybrid, and generative AI.

References

[1] ALTAI, "Assessment list for trustworthy artificial intelligence (altai)", High-Level Expert Group on Artificial Intelligence, European Commission, Tech. Rep., 2019.

[2] A. Awadid et al., "AI systems trustworthiness assessment: State of the art", in *Workshop on Model-based System Engineering and AI, 12th International Conference on Model-Based Software and Systems Engineering (Modelsward)*, 2024.

[3] K. Quintero et al., "An end-to-end method for operationalizing trustwothiness in AI-based critical systems", in *15th Int. Conf. on Performance, Safety and Robustness in Complex Systems and Applications*, 2025.

[4] European Union Aviation Safety Agency, "Artificial Intelligence Roadmap: A Human-Centric Approach to AI in Aviation", Tech. Rep., 2023.

[5] J. Mattioli et al., "AI engineering to deploy reliable AI in industry", in *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, IEEE, 2023, pp. 228–231.

[6] F. Kaakai and P. Raffi, "Towards multi-timescale online monitoring of AI models: Principles and preliminary results", in *SafeAI, AAAI's Workshop on Artificial Intelligence Safety*, vol. 3381, 2023.

[7] M. Gonzalez et al., "Introducing RUM: A Methodological Contribution for Engineering Trustworthy AI Components in Industrial Systems", in *Proceedings of the AAAI Symposium Series*, AAAI, vol. 7, 2025, pp. 153–160.

[8] X. Li et al., "From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V", *IEEE Intelligent Systems*, vol. 37, no. 4, pp. 18–26, 2022.

[9] K. Kapusta et al., "Protecting ownership rights of ml models using watermarking in the light of adversarial attacks", *AI and Ethics*, vol. 4, no. 1, pp. 95–103, 2024.

[10] G. Soudain, "EASA Artificial Intelligence (AI) Concept Paper Issue 2: Guidance for Level 1&2 machine learning applications - Issue 2", European Union Aviation Safety Agency, Tech. Rep., 2024.

[11] J. Mattioli et al., "An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering", *AI and Ethics*, vol. 4, no. 1, pp. 15–25, 2024.

[12] J. Machrouh et al., "Qualification/validation of AI-augmented ATM solutions for sustainable aviation", *Towards Sustainable Aviation Summit*, 2025.