Evaluation of Robustness, Reliability, and Safety of an AI Based System

Lucas Mattioli, Annia Abtout, Martin Gonzalez, Afef Awadid, Kevin Mantissa, Faouzi Adjed, Joseph Machrouh, Jaime De Oliveira, Christophe Guettier, Hatem Hajri, Juliette Mattioli

www.thalesgroup.com

IARIA

SystemX INSTITUTE OF RESEARCH AND TECHNOLOGY

TALAN Positive innovation

THALES Building a future we can all trust

SAFRAN

cortAIx Artificial Intelligence for Mission Critical Systems

> **Juliette Mattioli** (PhD), started as a high school mathematics teacher in 1982.

> She joined the Thomson-CSF research team in 1990 as a PhD student working on **pattern recognition** systems, and in 1993 earned a PhD in mathematics applied to AI from Paris IX Dauphine University, with a thesis on "Inverse Problems and Differential Relations in Mathematical Morphology".

> In 1996, she turned to **decision support, combinatorial problem-solving**, and **AI planning.**

> In 2001, she directed an Industrial Research Lab focusing on combinatorial optimization, AI planning, constrained programming, and decision support.

> In 2005, she established a new research focus **on information fusion combining semantics and knowledge-based reasoning,** and she led research at Thales focusing on information fusion, combinatorial problem-solving, optimization and decision support, cyber security and big data.

> She joined Thales's Technical Directorate in 2010. In 2017, she was appointed Senior AI Expert, and in 2024 **Thales AI Fellow.**

> In 2017, she was a G7 Innovators' Conference on artificial intelligence representative from France, as a member of the # FranceIA mission

> She received in 2026 a "Margaret Award" for intrapreneurship for her work on **trustworthy AI engineering.**

Agenda

01

Rationale & Regulation Landscape

02

Core Definitions & AI Paradigms

03

AI-based critical system lifecycle

04

Paradigm-Specific Trustworthy Assessment Methods

05

Open Challenges & Research Directions

06

Wrap-up and Q&A

Rationale

- AI is now deployed in high-stakes domains: aerospace, healthcare, automotive, defense
- Failures in these domains can have severe, irreversible consequences
- Major challenges remain in ensuring trustworthy AI.
- Need for paradigm-aware framework for evaluating AI robustness, reliability, and safety.

Regulatory Landscape

> EU AI Act (2024)

Legally binding for high-risk AI systems

> EASA AI Roadmap 2.0

Certification framework for aviation AI

> ISO/IEC 5338

AI-specific software lifecycle processes

> DO-178C / ARP 6983

Avionics software & AI system standards

RRS: Core Definitions

ROBUSTNESS

Does the system maintain performance under perturbations?

Extends beyond adversarial attacks to distributional shift — where deployment data differs from training data. Models must adapt to varying contexts, not just resist noise.

Evidence:

Adversarial test sets · OOD benchmarks · Stress tests

Standards: ISO/IEC 24029 · EASA DAL requirements

RELIABILITY

Does the system produce correct outputs consistently?

Measured via failure rate, error rate, confidence calibration, and temporal stability. A model scoring well on a static test set may still cause errors if miscalibrated.

Evidence:

Performance metrics · Failure rate analysis · Uncertainty quantification

Standards: DO-178C · ISO 26262 · IEC 61508

SAFETY

Does the system avoid causing harm?

Broader socio-technical goal. Reliability is necessary but insufficient. A system can function reliably yet pose unacceptable risks if objectives conflict with safety constraints.

Evidence:

Hazard analysis · FMEA/FMEDA · Runtime monitoring logs

Standards: ARP 4754B · ARP 4761A · MIL-STD-882

Interdependencies of RRS

Reliability alone does not guarantee Safety

An LLM may reliably generate fluent text while producing harmful or misaligned content — objectives matter.

Robustness is a prerequisite for Reliability

A system withstanding adversarial inputs and distribution shifts will produce more consistent outputs.

Neglecting any one dimension creates technically capable but untrustworthy systems → all three must be assessed jointly.

Four AI Paradigms & Dominant Failure Modes

Data-Driven AI

Examples: Deep neural networks, SVMs, evolutionary algorithms

⚠ **Adversarial attacks · Distributional shift · Bias**

Performs well in controlled settings; fragile to real-world unpredictability

Symbolic AI

Examples: Expert systems, logic programming, constraint solvers

⚠ **Incomplete rule bases · Logical inconsistencies**

Strong formal guarantees; effectiveness hinges on knowledge-base quality

Hybrid AI

Examples: Neuro-symbolic, Physics/Geometry-Informed NNs

⚠ **Interface mismatches between constituent components**

Combines strengths; symbolic constraints can enhance trustworthiness

Generative AI

Examples: LLMs, diffusion models, foundation models

⚠ **Hallucinations · Prompt injection · Bias · Misalignment**

Scale and emergent behaviour undermine traditional verification methods

Industry Examples: Aerospace & Healthcare

Predictive Maintenance
Must remain accurate with incomplete/biased historical data — supports reliable maintenance decisions (Safran)

Confidence Miscalibration
Temporal drift in populations and imaging protocols causes systematic yet hidden errors

AI-Driven Flight Optimizers
Fail-safe mechanisms for edge cases — severe weather, conflicting ATC instructions (Thales TopSKy)

Clinician Dependence
A diagnostic AI may excel in trials yet be unsafe if clinicians over-rely or lack human-override procedures

Pilot Assistance (LARD)
Continuous monitoring for 'black box' decision drift — formal methods or stress testing essential (Airbus)

GenAI in Clinical Support
Fluent but wrong answers — no clear failure signals — clinically-tailored red-teaming essential

Trustworthiness in AI-based critical system impacts the overall engineering lifecycle

Needs for engineering methods and tools to support the overall lifecycle of critical AI-based systems…

… To ensure qualification and compliance with regulations and standards

Trustworthiness in AI-based critical system impacts the overall engineering lifecycle

From requirements & specifications

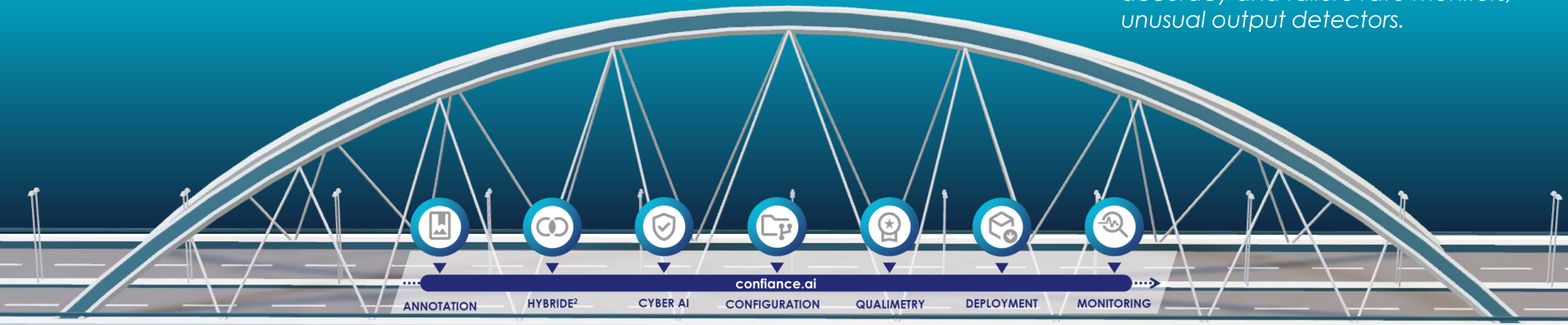
→ Stakeholder reqs.
→ Sys. & AI/ML specs.
→ Sys & AI/ML archis

ML/AIOps

To system deployment & maintenance

→ Monitoring
→ Toward qualification and certification

Continuous assurance, drift monitors, accuracy and failure rate monitors, unusual output detectors.



AI engineering: a necessarily condition to deploy trustworthy AI

Operational Design Domain (ODD)

Validation

Verification

Operational Analysis regarding Intended Purpose

To ensure qualification and compliance with regulations (e.g. AI Act) **and standards** (e.g. ARP6983 in aeronautics)

Test plan coverage, independence of V&V, learning process verification, formal analysis.

Integration into the AI System Lifecycle (ISO/IEC 5338)

Inception	Design & Development	Verification & Validation	Deployment	Operation & Monitoring
Define measurable RRS requirements Specify Operational Domain (ODD) Link to system hazard analysis Performance & robustness targets	Separation of concerns architecture Explicit uncertainty representation Redundancy & diversity strategies Graceful degradation modes	OD-covering test plan Independent V&V team Formal methods (where feasible) Learning process verification	Transition from test to operational Deployment configuration freeze Shadow-mode validation for LLMs Change impact assessment	Drift detection monitors Accuracy & failure rate tracking Anomaly / unusual output detection Continuous re-evaluation triggers

Trustworthiness cannot be added after the fact — it must be engineered at every phase.

Key Engineering Concepts: Intended Purpose, ODD & V&V

> Operational Design Domain (ODD)

- Defines the conditions under which the AI constituent is designed to function as intended
- Sets reference domain for all trustworthiness attributes
- Must be complete, consistent and human-readable
- Tests must verify performance throughout Operational Domain (OD) — inputs outside must trigger defined safe behaviors
- Deviating from intended purpose = misuse; operating outside OD = out-of-distribution risk

> V&V Principles

- **Test Coverage:** All OD scenario classes sampled; safety-relevant minority classes over-sampled
- **Independence:** V&V team independent of development — must not have contributed to training data
- **Process V&V:** Verify dataset quality, training pipeline reproducibility, hyperparameter selection
- **Formal Methods:** Formal proof for symbolic parts + statistical for neural → system-level guarantees
- **Post-Deployment:** Continuous drift monitoring, performance threshold triggers, change impact assessment

Operational Domain is both a design constraint and an assessment tool

Paradigm-Specific Trustworthy Assessment Methods

Data-Driven AI

Robustness	Reliability	Safety	Key Gap
Adversarial testing · Robustness certification · OOD benchmarks	Cross-validation · Calibration metrics · Temporal drift detection	STPA-adapted hazard analysis · Statistical (not causal) methods	Methods are statistical — failure explanation remains limited

Generative AI

Hybrid AI

Symbolic AI

Data-Driven AI

Paradigm-Specific Trustworthy Assessment Methods

Symbolic AI

Robustness	Reliability	Safety	Key Gap
Formal verification avoids adversarial vulnerability — deterministic reasoning	Model checking · Theorem proving · Correctness guarantees	Traceable hazard paths · Depends on knowledge-base completeness	Rigid reasoning — maintaining complete, consistent KB is core challenge

Generative AI

Hybrid AI

Symbolic AI

Data-Driven AI

Paradigm-Specific Trustworthy Assessment Methods

Hybrid AI

Robustness	Reliability	Safety	Key Gap
Validating symbolic interpretation of data-driven outputs · Physics constraints	Statistical testing + formal verification combined for system-level guarantees	Contract-based design · Compositional verification	Symbolic constraints restrict outputs to logically valid conclusions

Generative AI

Hybrid AI

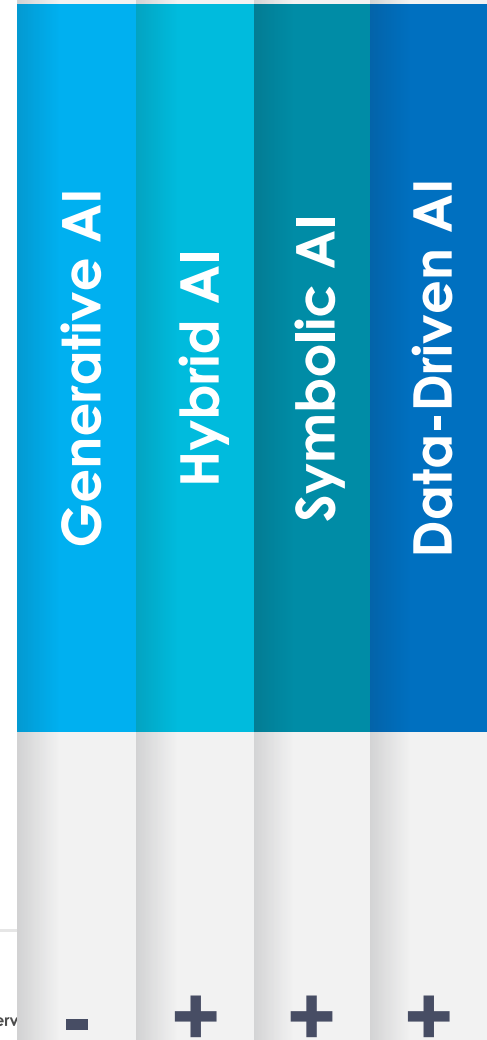
Symbolic AI

Data-Driven AI

Paradigm-Specific Trustworthy Assessment Methods

Generative AI

Robustness	Reliability	Safety	Key Gap
Prompt robustness testing · Jailbreak detection · Watermarking	Factuality metrics · Bias audits · Truthfulness assessment	Red-teaming · RLHF · Alignment protocols (no formal guarantees)	Emergent behaviour — traditional verification is fundamentally inadequate



Open Challenges & Research Directions

1 Scalable Formal Verification

Formal methods offer strongest guarantees but not yet applicable to large NNs.
Key direction: compositional methods, abstraction refinement, architectures designed for verifiability.

2 ODD Specification for High-Dimensional Inputs

For NLP, raw sensor data, or visual scenes, defining ODD precisely is an open problem.
Need: formal specification languages + automated boundary detection tools.

3 GenAI Safety Verification

Hallucination and prompt injection lack systematic verification. Manual red-teaming is costly, incomplete, and non-reproducible.
Automatic red-teaming and formal behavioral specs are open problems.

4 Causal Models for Assurance

Current assessment uses statistics to measure what, not why. Causal models could reveal root causes, predict failure conditions, and greatly strengthen safety cases.

5 Evidence Aggregation

Multiple assessment dimensions need formal methods for combining heterogeneous evidence — uncertainty propagation in safety cases and formal assurance case logics.

Future work extends the approach beyond data-driven AI — covering symbolic, hybrid, and generative AI engineering methods and tools.

Wrap-up

1. **Assessment must be paradigm-aware — methods for symbolic solvers, neural networks, and LLMs differ; a uniform approach leaves dangerous gaps.**

2. **Robustness, Reliability, and Safety are interdependent but distinct — each requires separate assessment before integration into a safety case.**

3. **Assurance is a lifecycle activity, not a one-off certification — changing environments, drift, and new adversarial techniques demand continual evidence.**

This work is a step toward an AI Assurance Engineering discipline that evolves with the technology it governs.

THALES
Building a future we can all trust

Thank you

www.thalesgroup.com