

Assuring AI/ML-Enabled Safety-Critical Software in the Space Domain

Alberto Petrucci

Thales Alenia Space

alberto.petrucci@thalesaleniaspace.com

GranSasso Science Institute

alberto.petrucci@gssi.it

September 24, 2025



Why AI/ML in Space?

Space missions demand unprecedented levels of **precision, autonomy, and reliability**. Traditional software validation methods, while mature, are insufficient for AI/ML systems due to their unique characteristics.

- AI/ML enables:
 - **Autonomous decision-making** in unpredictable environments.
 - **Real-time processing** of large and complex datasets.
 - **Reduced reliance** on ground-based human intervention.
- **Key Challenge:** There are no universally accepted standards for validating safety-critical AI/ML systems in space.

ECSS Standards Framework

The European Cooperation for Space Standardization (ECSS) provides a robust framework for software qualification; however, it was designed primarily for traditional systems.

- **ECSS-Q-ST-80C**: Defines software product assurance requirements.
- **ECSS-E-ST-40**: Covers software engineering lifecycle processes.
- **ECSS-E-HB-40-02A**: Guidelines for AI/ML development, verification and validation.

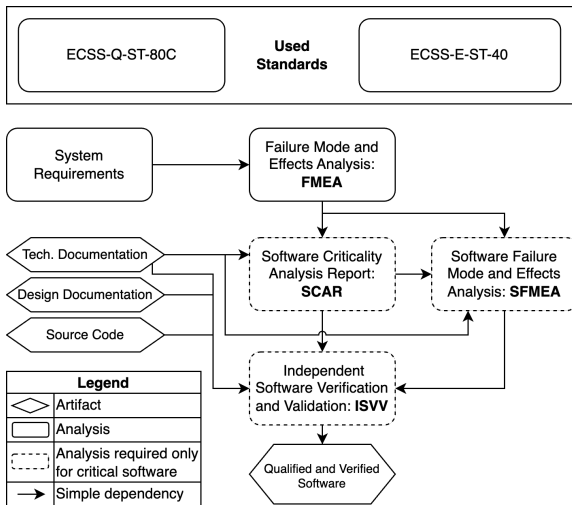
Software Criticality Levels

Space software is categorized by the severity of potential failures, which determines the required rigor of independent validation.

Level	Risk Description	ISVV Requirement
CAT-A	Life-threatening or mission-critical failure	Level 2: Deep code & design analysis
CAT-B	Major mission degradation or hardware damage	Level 1: External analysis
CAT-C	Recoverable issue or minor data corruption	Standard development & testing

Table 1: Criticality levels and corresponding validation rigor

Traditional Qualification Process



Testing Paradigm Shift

Software Development	Software Testing	ISVV
Hand Written Model Based Auto generated	Unit Testing Functionality Testing Integration Testing	Technical Analysis Review Design Analysis Review Code Analysis Review
AI/ML Development	AI/ML Testing	AI/ML Validation
Different modeling tools/frameworks (TensorFlow, PyTorch, ...) Different training techniques (quantization aware...) Different approaches for different cases (DNN, RNN, CNN, ...)	Performance testing Robustness testing Corner Case testing Single Event Upsets testing	Technical Analysis Review Design Analysis Review Training Dataset Biases Review Explainability Analysis Review Monitorability Analysis Review Provability Analysis Review Robustness Analysis Review SEUs Analysis Review
(A)	(B)	(C)

ISVV Adaptation Challenges

Traditional ISVV does not fully transfer to AI/ML — requiring new validation methods.

Traditional ISVV Activities

- **TAR:** Requirements validation
- **DAR:** Design evaluation (partial)
- **CAR:** Code-level analysis

AI/ML Validation Needs

- Dataset analysis (bias & coverage)
- Explainability/interpretability
- Robustness testing (adversarial inputs)
- Resilience to SEUs (radiation faults)

Qualifying AI/ML for CAT-C/B/A

The approach to qualification must be tailored to the system criticality level, ensuring that the appropriate rigor is applied to mitigate risks effectively.

Crit.	Qualification Solutions
C	Basic testing and documentation to ensure functionality and traceability .
B	Dataset V&V to ensure representativeness and quality, combined with Independent Model Verification and Validation for thorough assessment.
A	Formal verification, redundancy, continuous monitoring , and provability analysis to ensure safety and reliability.

Motivation: Trustworthy AI in Safety-Critical Aerospace

The integration of AI/ML in aerospace offers transformative benefits but also introduces significant risks that must be managed.

- **Benefits:** Improved autonomy, efficiency, and mission success rates.
- **Risks:** Catastrophic failures due to unpredictable AI/ML behavior.
- **Regulatory Landscape:** Emerging standards (e.g., EU AI Act, ISO 21448) aim to address these risks but remain fragmented.

Key Contributions

This work provides a comprehensive framework for adapting existing standards and introducing new methodologies for AI/ML assurance.

- **Gap Analysis:** Identifies limitations of current ECSS standards for AI/ML.
- **Aerosafe Methodology:** Manages residual uncertainty in AI/ML systems.
- **AMLAS Integration:** Aligns Assurance of Machine Learning in Autonomous Systems (AMLAS) with ECSS processes.
- **Practical Tools:** Checklists and lifecycle-phase guidance tailored to system criticality.
- **Assurance Confidence:** Structured safety arguments with explicit confidence levels.

AMLAS Overview

AMLAS provides a lifecycle-oriented approach to AI/ML assurance, focusing on uncertainty management and safety case construction.

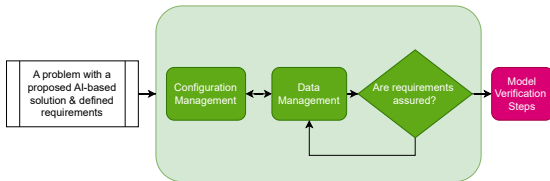
- Structured methodology for AI/ML assurance.
- Addresses lifecycle uncertainty, robustness, and explainability.
- **Does not fully align** with ECSS Verification, Validation, and Review processes.

Aerosafe Methodology

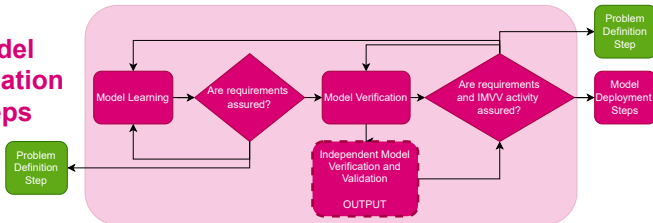
The **Aerosafe methodology** extends AMLAS to provide a tailored solution for aerospace applications, ensuring compliance with ECSS standards.

- Defines assurance scope and allocates safety requirements.
- Manages ML-relevant data across the lifecycle (quality, bias, configuration).
- Embeds learning, verification, and deployment activities within ECSS reviews.
- Introduces Independent Model Verification and Validation (IMVV) for high-criticality systems.

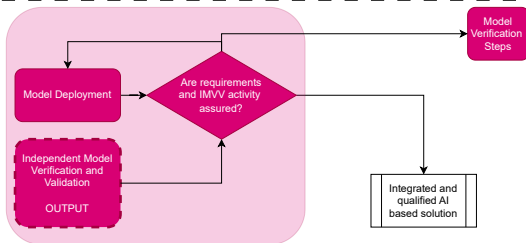
Problem Definition Steps



Model Verification Steps



Model Deployment Steps



- Activity for all Criticality
- Activity for Criticality B & A

Independent Model Verification and Validation (IMVV)

IMVV provides an independent and rigorous assessment of AI/ML models, ensuring they meet safety and performance requirements.

- **Core Activities:**
 - Validation of requirements and data integrity.
 - Assessment of model accuracy and deployment risks.
- **Category A Systems:** Additional regulatory conformance checks and post-deployment monitoring.

Assurance Artifacts and Checklists

Aerosafe delivers comprehensive and traceable artifacts to support safety assurance and certification.

- Data specifications and versioned datasets.
- Model documentation and simulation test specifications.
- Safety case modules with explicit confidence levels.

Discussion and Limitations

While the **Aerosafe methodology** addresses many gaps, several challenges remain in the assurance of AI/ML systems.

- **Distributional Shifts:** Changes in data distribution over time can impact model performance.
- **Rare Edge Cases:** Infrequent but critical scenarios must be identified and addressed.
- **Explainability:** Understanding AI/ML decision-making remains a challenge.
- **Scalability:** Formal verification and high-fidelity simulation are resource-intensive.

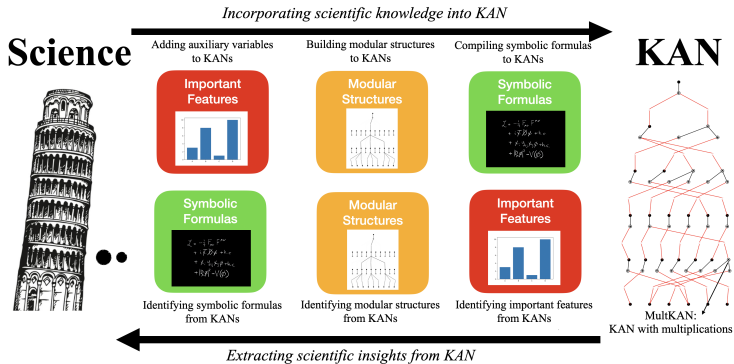
Challenge: AI + Science Synergy

AI/ML and scientific discovery operate on different paradigms, creating both challenges and opportunities for synergy.

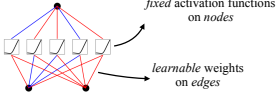
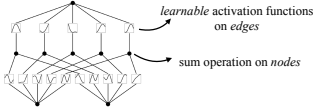
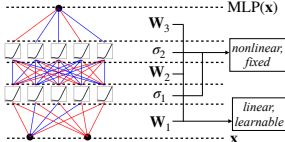
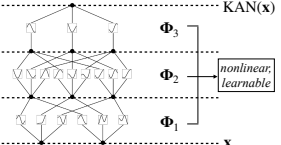
- **Inherent Incompatibility:**
 - AI is **connectionist** (e.g., neural networks, data-driven).
 - Science is **symbolic** (e.g., theories, equations).
- **Goal:** Develop AI/ML systems that support curiosity-driven scientific discovery.

KANs for Science

KANs offer a promising approach to bridge the gap between AI/ML and scientific discovery by combining interpretability with performance.



KAN vs. Traditional Multi-Layer Perceptron (MLP)

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{M(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	(a)  fixed activation functions on nodes learnable weights on edges	(b)  learnable activation functions on edges sum operation on nodes
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	(c)  MLP(x) \mathbf{W}_3 σ_2 \mathbf{W}_2 σ_1 \mathbf{W}_1 \mathbf{x} nonlinear, fixed linear, learnable	(d)  KAN(x) Φ_3 Φ_2 Φ_1 \mathbf{x} nonlinear, learnable

Conclusion: KANs for Safety-Critical Applications

The **Aerosafe methodology** provides a pragmatic and robust framework for the safe integration of AI/ML in aerospace systems.

- Integrates AMLAS with established ECSS standards.
- Manages uncertainty and tailors assurance activities to system criticality.
- Emphasizes data management, IMVV, and formal reviews.

KANs demonstrate significant potential for safety-critical applications due to their interpretability and performance.

- Better interpretability compared to traditional deep neural networks (DNNs).
- Comparable or improved accuracy in many domains.
- Suitable for applications where transparency and symbolic reasoning are essential.

Thank You!

We welcome your questions and feedback.