Zurich University of Applied Sciences

certAInty – A Certification Scheme for AI Trustworthiness An Innosuisse Project

Centre for Artificial Intelligence (CAI)[1]
Institute of Applied Mathematics and Physics (IAMP)[2]
CertX AG

Philipp Denzel[1], Oliver Forster[1], Yann Billeter[1],, Frank-Peter Schilling[1], Ricardo Chavarriaga[1], Carmen Mei-Ling Frischknecht-Gruber [2], Stefan Brunner [2], Monika Reif [2], Joanna Weng [2]

Background

The European Union's AI Act, effective August 1, 2024, establishes a comprehensive regulatory framework for AI systems. As organisations and certifiers prepare to demonstrate compliance, practical guidelines for achieving AI trustworthiness remain limited, both within the EU and globally. The AI Act categorises AI systems by risk – **unacceptable, high, limited, and minimal** – while unacceptable-risk practices, such as biometric classification, emotion recognition, subliminal manipulation, and social scoring, are prohibited. High-risk systems, particularly in sectors like healthcare, transport, education, and energy, must meet strict requirements for market entry, including CE-marking certification.

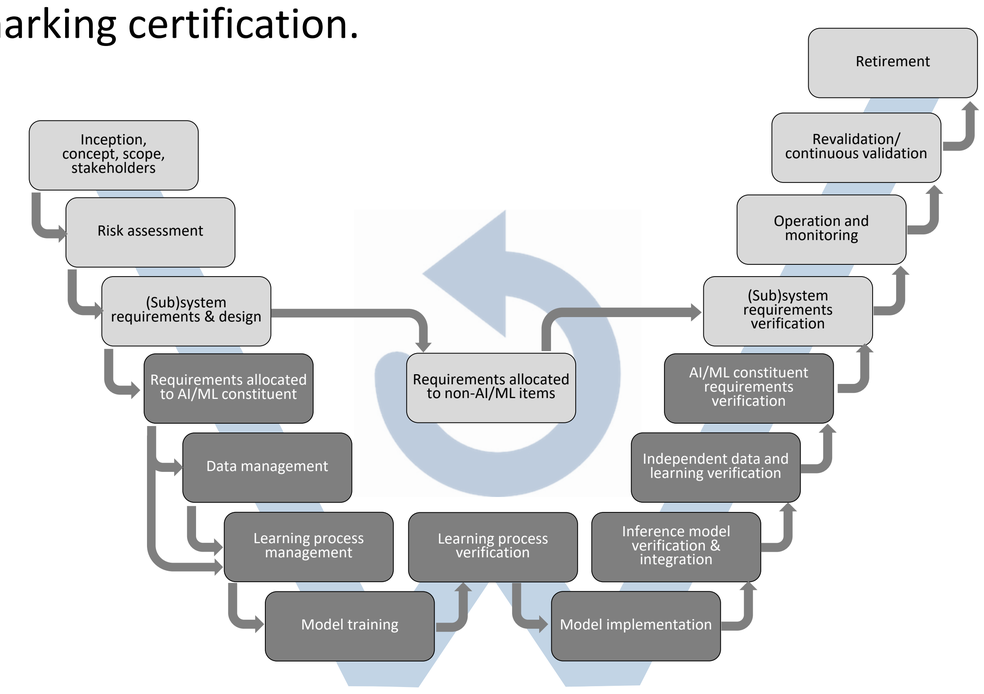

Figure 1 – The AI-based system lifecycle (source: EASA adapted)

High-risk AI systems also undergo conformity assessments and third-party certification to ensure transparency, human oversight, fairness, reliability and safety and security, as well as the protection of fundamental rights. Meanwhile, limited-risk AI systems are obligated to maintain the key aspect of transparency regarding their purpose, function, and decision-making processes. Supporting these regulations are standards from organisations such as ISO/IEC, IEEE, and NIST. ISO/IEC standards address AI terminology, performance metrics, data quality, and ethics. Whereas the IEEE P7000 series focuses on the ethical implications of AI technologies, while NIST provides frameworks for risk management, data quality, and transparency, offering essential guidance for demonstrating compliance and promoting best practices.

CertAInty Project Overview

The CertAInty project has developed a comprehensive Certification Scheme for AI Systems to ensure their trustworthiness. The scheme defines clear objectives and corresponding means of compliance, criteria, and evaluation measures aligned with relevant standards and regulations. It offers technical and scientific methods for assessing key AI properties, such as transparency, human oversight, data governance, reliability, fairness, and safety and security.

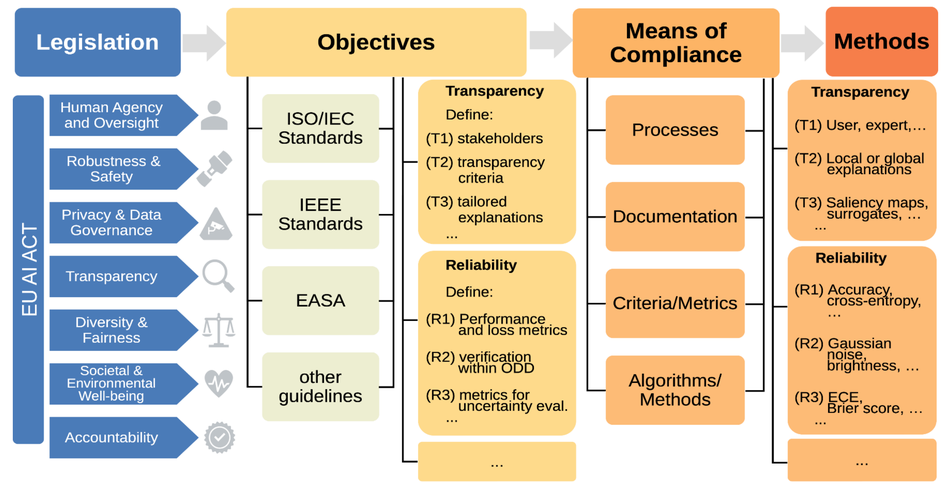

Figure 2 – Certification scheme for AI systems. It enables identification of application-specific requirements, respective means of compliance and technical methods for conformity assessment

These methods support systematic evaluations throughout the AI lifecycle — from risk assessment, requirements and data acquisition to development, testing, and deployment. Based on an analysis of 38 standards (ISO/IEC, IEEE, EASA) and integration of EU legislation, the scheme bridges regulatory requirements with technical assessments. It identifies 95 techniques for measuring attributes like explainability, robustness, and security, ensuring compliance with the EU AI Act.

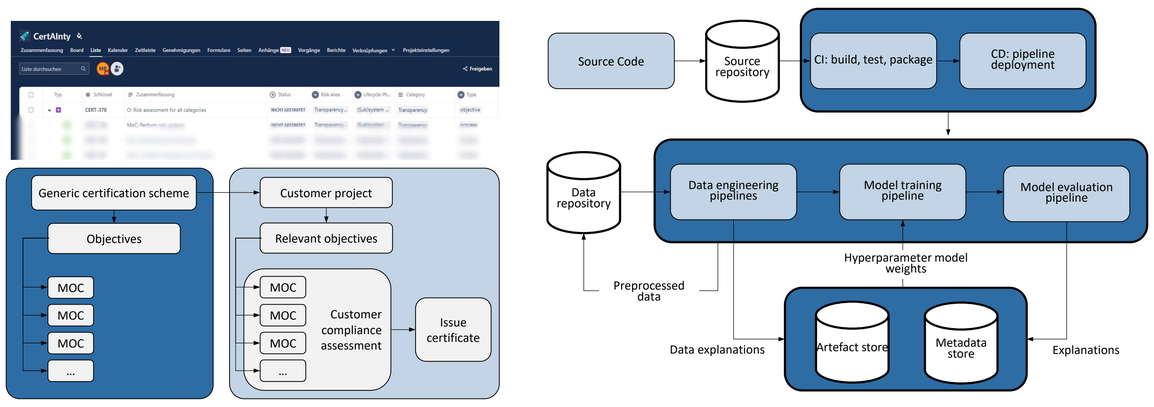

Figure 3 – Practical application: On the left a visualisation of the certification workflow based on JIRA and on the right an overview of the MLOps system architecture.

To ensure a reliable and practical approach to AI certification, the CertAInty project uses Jira for requirements management. This tool ensures traceability, systematic documentation and real-time collaboration in line with standards such as ISO 9001, ISO/IEC 27001 and ISO/IEC 23894. Jira's features including version control, decision tracking and customisable dashboards improve efficiency and consistency throughout the certification process. Complementary MLOps practices ensure trustworthiness by design, incorporating best practices at every stage of the AI lifecycle. Our MLOps infrastructure supports model development, training, continuous integration and continuous deployment and monitoring, using tools such as GitHub, MLflow and Apache Airflow. This setup maintains traceability, automates testing and ensures ongoing standards compliance. The workflow has been validated in three high-risk use cases - medical applications, autonomous driving and vehicle detection at construction sites - aligning with the requirements of the EU AI Act and demonstrating effective certification processes for different data types and compliance requirements.