AI Assessment in Practice: Implementing a Certification Scheme for AI Trustworthiness

Carmen Frischknecht-Gruber
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Philipp Denzel
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Monika Reif
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Yann Billeter
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Stefan Brunner
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Oliver Forster
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Frank-Peter Schilling
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Joanna Weng
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Ricardo Chavarriaga
Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland

Abstract

The trustworthiness of artificial intelligence systems is crucial for their widespread adoption and for avoiding negative impacts on society and the environment. This paper focuses on implementing a comprehensive certification scheme developed through a collaborative academic-industry project. The scheme provides practical guidelines for assessing and certifying the trustworthiness of AI-based systems. The implementation of the scheme leverages aspects from Machine Learning Operations and the requirements management tool Jira to ensure continuous compliance and efficient lifecycle management. The integration of various high-level frameworks, scientific methods, and metrics supports the systematic evaluation of key aspects of trustworthiness, such as reliability, transparency, safety and security, and human oversight. These methods and metrics were tested and assessed on real-world use cases to dependably verify means of compliance with regulatory requirements and evaluate criteria and detailed objectives for each of these key aspects. Thus, this certification framework bridges the gap between ethical guidelines and practical application, ensuring the safe and effective deployment of AI technologies.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Social and professional topics → Computing / technology policy; Information systems → Information systems applications

Keywords and phrases AI Assessment, Certification Scheme, Artificial Intelligence, Trustworthiness of AI systems, AI Standards, AI Safety

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.1

Funding This work was co-financed by Innosuisse (101.650 IP-ICT).The contribution of R.C. was partially funded by the Wellcome Trust [Grant number: 226486/Z/22/Z].

Acknowledgements We would like to acknowledge the support and collaboration of CertX AG in the development of the certification scheme discussed in this paper.

1 Introduction

Global efforts are underway to implement frameworks for assessing and regulating artificial intelligence (AI) systems. The most imminent of these efforts is the EU Artificial Intelligence

Act [8]. The AI act gradually comes into force starting 1 August 2024, which means organisations and certifiers are in dire need of building their capacity to prove and assess compliance now. However, despite this and other forthcoming regulations around the globe, there remains a significant lack of practical guidelines and methodologies for both achieving and assessing the trustworthiness of AI-based systems (AIS). Although there has been extensive development of ethical guidelines for AI, c.f., Jobin et al. [20], the practical application of these principles remains vague. The lack of specificity in the operationalisation of these guidelines presents a challenge to their effective implementation across various AIS. The introduction and deployment of inadequately understood and unreliable AI technologies can result in significant societal harm. These include the exclusion or discrimination of minorities due to inherent biases and even physical injuries resulting from erroneous decision-making by AIS, such as in human-robot interactions or misdiagnoses in the healthcare sector. Furthermore, such technologies have the potential to exacerbate existing educational disparities, lead to unfair legal outcomes, and increase inequality. There is also a substantial risk of environmental damage, privacy breaches, and cybersecurity vulnerabilities. It is, therefore, imperative to develop tools that allow for AIS to be thoroughly vetted for responsibility and ethical considerations to mitigate these risks and protect societal well-being.

To address this issue, the authors, in collaboration with a certification company, are developing a certification scheme for AIS. This scheme is intended as a practical guide and provides corresponding tools for developers and regulators to evaluate and certify the trustworthiness of AIS throughout their lifecycle, including requirements, data acquisition, model development, testing, deployment, and operation. It builds upon current standards and guidelines of a number of bodies, including ISO/IEC, IEEE, EASA, as well as other guidance documents [19, 18, 12, 32, 27], in addition to EU legislation. A total of 38 documents were subjected to analysis, and the objectives and the various means of complying with them were derived from these inputs.

This certification scheme effectively bridges the gap between regulatory requirements, technical standards, and the specific scientific and technical methods needed to assess the properties of machine learning (ML) models. Noteworthily, regulatory requirements and technical standards do not provide clear instructions on which methods and metrics can be used to assess the properties for trustworthy AIS. To fill this gap, we evaluated and identified 95 technical methods for assessing the transparency, explainability, reliability, robustness, safety, and security of AI models. By doing so, the certification scheme complements existing approaches in trustworthy AI certification by incorporating cutting-edge research from the AI community on algorithmic techniques for determining and evaluating relevant model properties. As a result, it provides a complete operational framework that links the regulatory requirements to measurable objectives and methods to assess compliance with the EU AI Act and supports regulations in other jurisdictions.

This paper outlines the implementation and application of the certification scheme, with a particular focus on detailing the tools, workflows, and methodologies used to ensure both comprehensive compliance and practical utility. Furthermore, it describes how these tools and methodologies relate to objectives for means of compliance and demonstrates our approach to assessing the given requirements.

Identifying, tracing, and documenting appropriate objectives, procedures, and technical methods for assessing compliance requires adequate supporting tools. We address these needs by implementing the certification scheme within the management platform Jira. This is complemented by an automatised pipeline that implements algorithmic methods for assessing the trustworthiness of AI models. This pipeline is implemented according to best practices in AI engineering and Machine Learning Operations (MLOps) principles.

In the remainder of this paper, we give an overview of the current state of AI standardisation and regulatory efforts, highlighting key initiatives and guidelines. In Section 3, we outline the certification scheme, detailing relevant regulatory requirements, criteria, and the methodology for certification.

Then, we describe the implementation process, including tools and frameworks used to verify compliance, and how these relate to the regulatory requirements (Section 4). Finally, we summarise our findings and offer a discussion on the implications and future developments in AI certification (Section 5).

2 Background

The deployment and scalability of AI assessment frameworks face several key challenges, particularly in balancing practical implementation with theoretical underpinnings. One of the main obstacles lies in the aggregation of risks associated with AI systems, including bias, transparency, security vulnerabilities and ethical considerations. Current frameworks often address individual risks in isolation, but aggregating these risks in a way that provides a holistic assessment is complex. Many frameworks still lack widely accepted methods for this aggregation, leading to inconsistencies across industries and sectors. A significant need for interdisciplinarity also poses a challenge in scaling AI assessment frameworks. Inputs from law, ethics and computer science must be combined to form a coherent assessment approach. Managing this complexity requires the integration of technical AI safety measures with broader societal values, which is often challenging to operationalise at scale [37]. In terms of approaches, the risk-based approach used in regulations such as the EU AI Act offers a promising method for scaling up. This regulation categorises AI systems according to the level of risk they pose, from low-risk applications such as spam filters to high-risk systems such as healthcare AI. The EU AI law imposes strict regulatory requirements on high-risk systems to ensure safety and accountability. Conversely, AI systems classified as low risk are subject to a more flexible regulatory framework. Although these systems are not subject to the same stringent requirements, they must still comply with transparency and user information obligations. This risk-based classification ensures that regulatory oversight is aligned with the potential impact of AI systems, thereby increasing overall regulatory effectiveness while facilitating innovation in lower-risk areas. [9].

2.1 Regulation and Standards

Currently, there are significant global efforts to establish regulatory frameworks for AI. The EU has assumed a pioneering position with the AI Act, which is designed to establish a comprehensive regulatory framework for AIS [8]. In a similar vein, the United States issued an executive order in October 2023 with the objective of developing new standards for safe, secure, and trustworthy AI [43].

Standards and guidelines play a pivotal role in supporting binding laws and regulations by documenting best practices and providing a foundation for demonstrating compliance and certification. A considerable number of national and international organisations are engaged in a range of initiatives aimed at fostering trust in AI through the issuance of standards and guidelines. The ISO/IEC standards [19] address a plethora of AI-related aspects, including terminology, performance metrics, data quality, ethics, and human-AI interaction. These standards are currently in place, with more anticipated in the future.

Similarly, the IEEE is developing a certification program with the objective of assessing the transparency, accountability, bias, and privacy of AI-related processes [17]. The IEEE P7000 series [18] addresses the ethical implications of AI technologies. Other national entities, such as the National Laboratory of Metrology and Testing's (LNE) AI certification program, have established objective criteria for trustworthy AIS, emphasising ethics, safety, transparency, and privacy [23]. The NIST framework [27] offers guidance on the management of risks, the assurance of data quality, and the promotion of transparency and accountability in AIS, with related principles also emphasised in the AI Risk Management Framework [28]. Moreover, DIN/DKE offers comprehensive standardisation recommendations across all AI domains, facilitating a unified language, principles for development and utilisation, and certification [11]. In the field of aviation, the European Union Aviation Safety Agency (EASA) has introduced comprehensive guidelines for the safe utilisation of ML systems [42]. These guidelines provide support to stakeholders in the aviation sector at each stage of the lifecycle of AIS, from the initial stages of development through to operational use. The Fraunhofer Institute has developed a guideline for the design of trustworthy AI systems [32]. The guideline employs a six-dimensional evaluation framework to assess the trustworthiness of AIS, encompassing fairness, autonomy and control, transparency, reliability, safety and security, and privacy. In contrast to other contributions, the Fraunhofer guideline incorporates both process-related measures and technical methods to enhance the evaluation of AIS.

2.2 Frameworks

Capturing, tracing, documenting, and systematically evaluating requirements throughout the lifecycle of an AIS is an essential factor in trustworthy AI and its certification.

There are various methods and tools for the filtering and management of requirements, from very basic text files or Excel sheets to dedicated frameworks such as Confluence, Jira, Doorstop, Polarion, IBM Doors, Azure DevOps, and many more [3, 2, 5, 40, 33, 26]. In practice, the simple solutions do not provide the necessary flexibility and overview of the complicated relations between requirements. On the other hand, comprehensive requirement management frameworks are flexible but often less intuitive in their use and relatively expensive. After investigating several tools, we chose **Jira** (in its basic version, free) as a requirement management tool for the certification of AIS. Jira is a project management and issue-tracking software developed by Atlassian. It helps teams plan, track, and manage work efficiently, offering features like customisable workflows, real-time reporting, and integration with numerous other tools, making it a versatile solution for agile project management.

An important operational approach to scaling is the integration of Machine Learning Operations (MLOps). The role of MLOps principles and best practices in AIS development and operation, as well as its assessment, is twofold: First, Billeter et al.[4] and others [24] have advanced the idea of MLOps as the enabler of trustworthy AI by design. This means that following MLOps guidelines and principles during design, development and operation of an AIS, will lead to increased trustworthiness of the AIS. These practices include version control, continuous integration and deployment (CI/CD), automated testing, and monitoring. Second, the assessment of the trustworthiness of AIS also requires comprehensive evaluations of many objectives and means of compliance (MOC) derived from these requirements. Therefore, concepts like following best practices in AI engineering and MLOps are indispensable not just during AIS development but also during its assessment.

2.3 Algorithmic Tools for Trustworthy AI

While in some aspects of the verification of AI trustworthiness it is necessary to rely on qualitative results, in particular for model explainability or robustness, automated evaluation workflows mostly involve algorithmic methods with quantifiable output. Therefore, it is crucial to integrate assessment toolboxes which implement various algorithms and metrics, or rely on interfaces which allow for manual qualitative evaluation. There are a number of comprehensive toolboxes which implement appropriate technical methods paired with metrics, often isolated to assess specific aspects of AI trustworthiness such as transparency, reliability, or safety. For data and model explainability, industry-developed frameworks are Microsoft's InterpretML [25], Seldon's Alibi toolbox [38], IBM's AIX360 toolbox [45], Sicara's tf-explain [39], or PyTorch's captum API [6]. Additionally, Quantus [16] is a relatively new and complementary explainability toolbox which implements a growing number of metrics and provides interfaces for other toolboxes such as captum or tf-explain. Toolboxes for testing the reliability, robustness, and safety of an AI model are, e.g., MIT's Responsible AI Toolbox [41], Seldon's Alibi-Detect [47], IBM's ART [44] and UQ360 [46] toolboxes. In particular, there are many more toolboxes which implement specific tests for adversarial robustness, such as RobustnessGym [15], CleverHans [31], or Foolbox [34].

It is worth noticing that these toolboxes have been developed in parallel and, to a large extent, disconnected from the regulatory and certification frameworks. Hence, there suitability for compliance assessment is not entirely clear. Our analysis presents a significant step towards the integration of advances on both areas.

3 Overview of the Certification Scheme

The developed certification scheme for AIS encompasses several principal key aspects of trustworthiness, such as human oversight, transparency, reliability (including robustness), safety and security [10] at the moment. So we cover with the actual version already some of the key aspects of the EU AI act, other aspects as described in 1. Each aspect is considered to ensure that AIS operate effectively, ethically, and safely across various applications.

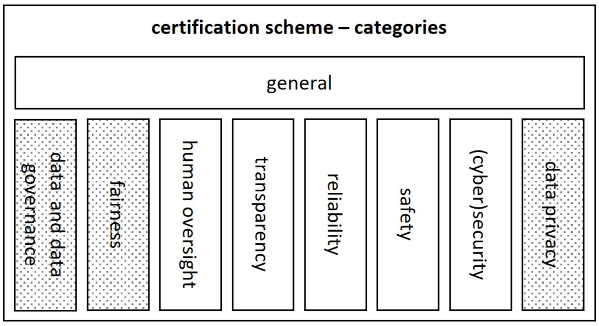

Figure 1 Extended key aspects of trustworthiness. The aspects of trustworthiness are as follows: data and data governance, fairness, human oversight, transparency, reliability, safety, (cyber)security, and data privacy. Within the certification scheme, the five non-shaded aspects are addressed, while the other three will be addressed at a later stage.

In addition, the certification scheme encompasses all relevant phases of the AIS lifecycle, as illustrated in Figure 2. The Certification Scheme employs a risk-based methodology in accordance with the EU AI Act. It commences with the concept of the system (including the role of the AI part within the overall system) and the associated system risk, subsequently progressing to the implementation of an AI model. The scheme culminates with the deployment, verification, validation, and operation of the system.

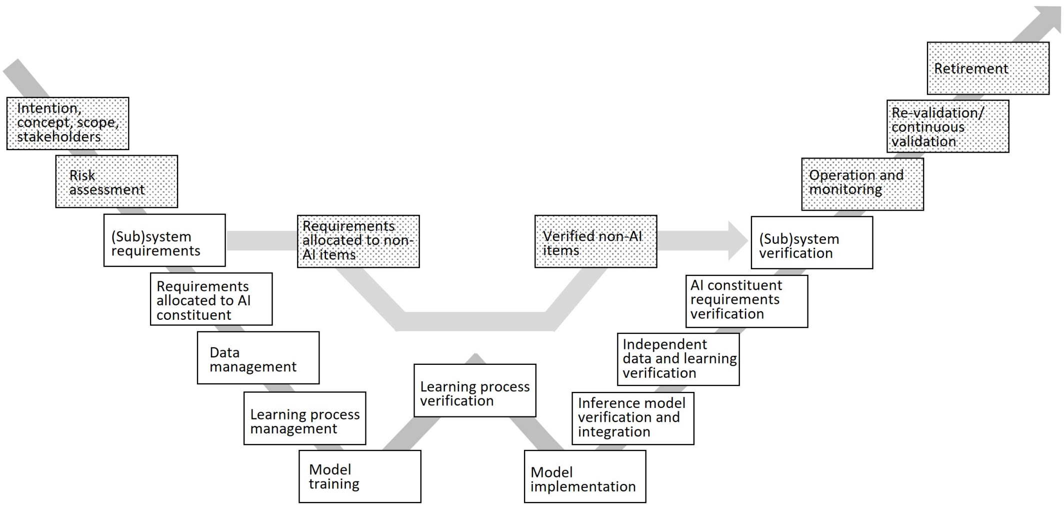

Figure 2 Illustration of the lifecycles encompassed by the certification scheme, including risk assessment, (sub-)system requirements and design, data management, learning process management, model training, verification steps, and operation and monitoring.

3.1 Key Aspects and Objectives

For each phase of the lifecycle, the scheme identifies and addresses the critical key aspects through the establishment of corresponding objectives. These build the basis for proving compliance with regulations as the EU AI Act and are derived from the EU AI Act, existing ISO standards [19] and the EASA guidances [42]. The objectives are refined according to different qualitative criteria and quantitative metrics (see Figure 3). Different MOCs have been defined to achieve compliance with the aforementioned objectives. One group of MOCs describes the process means that must be in place for a thorough development, verification, or management process. Others describe the documentation means to cover, for example, auditability and other record-keeping aspects. The last group of MOCs define the technical methods that must be applied to achieve compliance with the objectives posed. These MOCs establish the link to the different technical methods of the second technical part of the certification scheme.

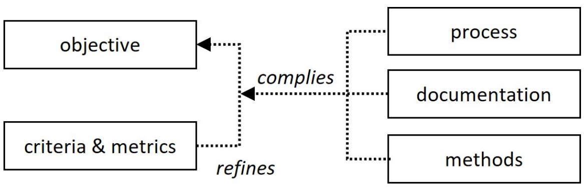

Figure 3 The interrelationship between objectives, criteria and metrics, and compliance methods is illustrated in the diagram. The left side depicts the objectives and their refinement through the application of criteria and metrics, while the right side shows the processes, documentation and methods that ensure compliance with these objectives and criteria.

Initially, the certification scheme focused on transparency and reliability, encompassing 29 and 44 objectives, respectively, with 100 and 156 MOCs. An updated version of the scheme additionally includes human oversight, safety and security alongside some general objectives relevant across multiple key aspects. Currently, the scheme covers:

- General Objectives: 5 objectives, 14 MOCs
- Human Oversight: 62 objectives, 65 MOCs
- Transparency: 29 objectives, 53 MOCs
- Reliability: 36 objectives, 105 MOCs
- Safety: 2 objectives, 6 MOCs
- (Cyber)Security: 5 objectives, 17 MOCs

The scheme includes a risk analyses and also addresses overlapping areas across key aspects, ensuring a comprehensive and integrated approach. Additional key aspects, such as data and data governance, will be implemented in the next step, and the key aspects of fairness and data privacy are planned for subsequent steps. In the following, we present two example objectives and their corresponding MOCs.

Objective 1: The applicant should define performance metrics to evaluate AIS performance and reliability.

- **MOC:** Define a suitable set of performance metrics for each high-level task to evaluate AIS performance and reliability.
- **MOC:** Define the expected performance with training, validation, and test data sets.
- **MOC:** Provide a comprehensive justification for the selection of metrics.

Objective 2: The applicant should identify and document the methods at AI/ML item and/or output level satisfying the specified AI explainability needs.

- **MOC:** Provide documentation of methods to provide explanations about the AI/ML item. The type and scope of the provided explanations should be chosen in terms of proportionality, considering the stakeholders.
- **MOC:** Specify the rules that apply to the current decision (e.g., for decision trees, list the selected branching next to the model output).
- **MOC:** Specify the most relevant attributes for a decision in linear regression models (e.g., for normalised inputs, the largest absolute coefficient value).
- **MOC:** For white-box models, use model-specific or model-agnostic methods for interpretability.

3.2 Key Aspects Overview

This section provides an overview of the key aspects covered by the scheme, including data governance, human oversight, transparency, reliability, and safety and (cyber)security.

3.2.1 Data and Data Governance

A dependable data set for a specific task requires careful attention to four key aspects: data quality, completeness, representativeness, and transparency. Data quality focuses on ensuring formal completeness and correctness and establishing reliability. The training, validation, and test data quality is assessed through qualitative and quantitative means (Figure 4). Correct annotations, task relevance, and data origin are crucial, alongside ensuring application coverage through metrics like class balance. Bias prevention requires unbiased training, validation, and test data, with fairness assessed via metrics like cosine similarity. Transparency ensures data is interpretable and preprocessing steps are clear, enabling verification by stakeholders.

(a) Data quality consists of six aspects. **(b)** Data coverage for the application.

Figure 4 Data quality (formal data completeness and correctness) and data coverage.

3.2.2 Human Oversight

Human oversight of AIS, also referred to as autonomy and control, addresses potential risks that may arise when autonomous AI components limit the ability of users or experts to perceive or act. This aspect of AI safety ensures that system autonomy is appropriately constrained when it deviates from normal operation. To assess human oversight, AIS are categorised into four levels based on human involvement [29]. The first level, Human Control (HC), involves the AI acting solely as an assistive tool, where humans are responsible for every decision and subsequent action based on the AI's output. At the Human-in-the-Loop (HIL) level, the AI operates partially autonomously but requires human intervention or confirmation, with humans monitoring and correcting its decisions as needed. The Human-on-the-Loop (HOL) level allows the AI to function almost autonomously, with limited human involvement for monitoring and occasional overrides. Finally, at the Human-out-of-the-Loop (HOOTL) level, the AI operates fully autonomously, handling tasks independently even in unexpected situations, with humans only involved in initial setup decisions like setting meta-commands in autonomous vehicles.

This key aspect includes objectives such as the implementation of human monitoring and control mechanisms, preservation of human decision-making capabilities, and ensuring the traceability of the AI component's decision-making process.

3.2.3 Transparency

Transparency in AI is essential to prevent potential harm and ensure systems are understandable to different stakeholders [36]. Transparency objectives are tailored to users, those affected (society), and experts (developers, providers, auditors and evaluators, authorities) (Figure 5). It involves setting criteria for interpretability and explainability, focusing on clarity, comprehensibility, and relevant metrics [7]. The interpretability of the ML model must be ensured through thorough documentation and visual aids like schematic diagrams. Explanation methods should be carefully chosen, justified, and documented, considering the audience's qualifications. These methods should be evaluated statistically and by human reviewers, with a system in place for addressing user queries. For experts, transparency also involves validating decisions, ensuring technical traceability, and maintaining reproducibility. Key considerations include the scope, design, and stability of explanation methods relative to model outputs.

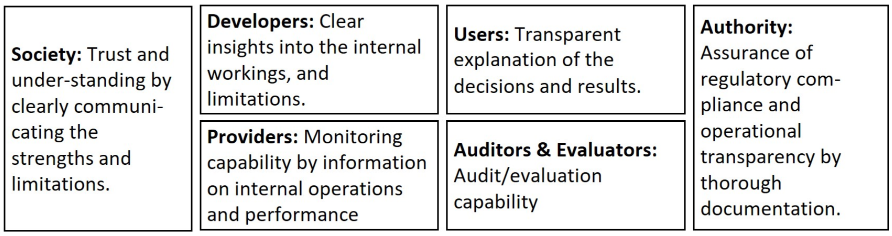

Figure 5 Transparency needs vary between stakeholders. The figure shows exemplary transparency requirements for some key stakeholders.

3.2.4 Reliability

Reliability in AIS is defined as the consistent execution of intended functions and also entails robustness, which pertains to maintaining performance under disturbances. An important concept is the Operational Design Domain (ODD), which delineates the specific conditions under which AIS can operate safely and effectively [35]. For developers, the ODD builds the basis for deriving detailed technical specifications that define the AIS input space, categorised into regular cases involving minor, expected disturbances; robustness cases where larger disturbances are encountered; and out-of-domain (OOD) cases, which involve data outside the application domain which may result in errors (Figure 6).

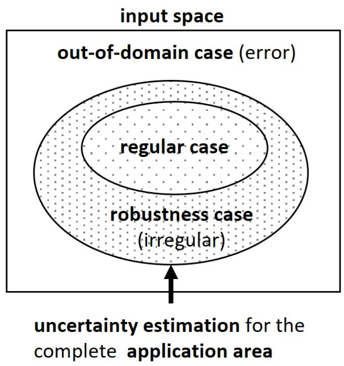

Figure 6 Visualisation of the input space divided into regular, robustness, and out-of-domain cases.

Consequently, reliability is assessed in the three input spaces, in addition to the estimation of uncertainty. The regular case ensures reliable performance through data coverage, augmentation, and performance metrics evaluation (see Figure 7). Robustness tackles challenging conditions by addressing vulnerabilities and adversarial attacks. In out-of-domain (OOD) cases, the focus is on catching errors and improving generalisation, while uncertainty estimation involves setting appropriate metrics, assessing both intrinsic and extrinsic uncertainties, and developing mitigation measures.

Additional process steps include evaluating model architecture, implementing optimisation techniques such as pruning or quantisation, ensuring reproducibility, conducting regular assessments, and meticulously documenting all activities.

In the certification scheme, reliability assessment involves over 55 metrics and 95 methods, with a subset of 35 metrics and 50 methods selected for empirical testing. This selection was based on relevance, execution time, reliance on available information, and computational costs.

Performance metrics

Regression
- (Mean) squared error
- (Mean) absolute error

Classification
- (Balanced) accuracy
- Micro/macro average
- (Balanced) F1-score
- Prevalence
- Precision
- False discovery/omission rate
- Positive/negative likelihood ratio
- True/False-negative rate
- True/False-positive rate
- Precision-recall curve
- Receiver operation characteristics (ROC)
- Lift
- Matthew's correlation coefficient
- Area under curve (AUC)
- Cohen's Kappa
- Confusion matrix
- Hinge loss

Computer vision
- Peak-signal-to-noise ratio (PSNR)
- Structural similarity
- (Mean) intersection over union (mIOU)

Clustering
- Silhouette value
- Adjusted mutual information score
- Completeness score

Ranking
- Mean reciprocal rank
- Discounted cumulative gain

Natural Language Processing
- Perplexity score
- BLEU score

Figure 7 List of performance metrics used for regression, classification, computer vision, clustering, ranking, and natural language processing.

Metrics vary across application domains and model objectives, so choosing the appropriate metric and method requires careful consideration of the model's goals, data characteristics, and desired outcomes. For example, formal verification employs logical and mathematical proofs to confirm system criteria, while model coverage analysis ensures comprehensive testing across various scenarios.

3.3 Safety and (Cyber)Security

The objective of safety is to minimise harm to people and the environment by designing AIS that incorporate corrective mechanisms for unexpected behaviours. This is of particular importance in contexts such as autonomous vehicles and healthcare, where errors can have significant and adverse consequences. (Cyber)security guarantees a system's integrity and availability by safeguarding it against unauthorised access, modification, or destruction. This encompasses the implementation of robust access controls, the assurance of data and model integrity, and the maintenance of system availability even in the event of an attack. Effective security measures are imperative for AIS in critical infrastructure, as breaches could result in significant damage. In order to enhance the security and resilience of AIS, adversarial training and verification are employed. Adversarial training is a method for enhancing the robustness of a model by exposing it to perturbations designed to deceive it, thereby identifying potential vulnerabilities.

4 Implementation of the Certification Scheme

The implementation and subsequent application of the AIS Certification Scheme to customers must meet established standards and regulations, requiring a carefully managed process. To achieve this, we evaluated several requirements management tools and ultimately selected Jira as the key tool for organising the objectives and associated means of compliance for our certification scheme. We then implemented an MLOps system based on state-of-the-art open-source tooling to perform the technical assessment of the AIS and evaluate the compliance with the defined objectives. In the following, we describe the requirement management system and the MLOps infrastructure.

4.1 Requirement Management Implementation

As written in section 2.2, Jira was chosen as requirement management tool to ensure traceability and effective management of the requirements. AI certification frameworks must adhere to internationally recognised standards, including ISO 9001 (Quality Management Systems), ISO/IEC 27001 (Information Security Management Systems), and the ISO/IEC 23894 (AI - Guidance on Risk Management). Such standards necessitate meticulous documentation, traceability, and periodic auditing to guarantee sustained compliance. The implementation of such requirements in a manual or disparate system would increase the risk of inconsistencies and errors, which would ultimately impact the efficiency and credibility of the certification process. It is therefore imperative that robust requirements management tools are employed.

Using Jira for requirement management ensures each objective and MOC is meticulously organised, facilitating clear communication and comprehensive oversight. Its ability to maintain detailed records and provide real-time updates is crucial for this task. Real-time collaboration and review capabilities are critical in aligning project tasks and reducing errors. The platform supports multi-user editing, allowing teams to work simultaneously from different locations. This live collaboration and features, such as decision tracking and impact analysis, ensure that the development of the certification scheme remains agile and responsive to changes. Additionally, the system's version control and history management provide a complete audit trail, which is crucial for maintaining consistency and verifiability.

Centralised management of objectives and MOCs in a digital environment allows for streamlined workflows and task alignment. We developed customised templates and dashboards for managing and tracing requirements. The possibility of sorting issues by attributes such as the tag COMPLETE was proven to facilitate requirement tracking in the evaluation we made of the platform. Each objective and MOC can be linked to others, showing relationships such as blocking issues and dependencies. The system's adaptability through the reusability of issues across different projects and its capacity for baseline creation significantly enhance the efficiency of the certification process. The platform facilitates organised and efficient project management by enabling tasks such as editing, organising decision-making, and managing tasks through a user-friendly interface. Integration with state-of-the-art tools, such as Git integration platforms like GitHub or GitLab, as well as business communication tools like Teams and Slack, along with the capability to create customisable pages, allows the tool to be precisely tailored to specific project needs.

The certification scheme is structured with a parent-child relationship between objectives and MOCs (Figure 8). Each issue type is defined by attributes, including description, main category, additional categories, lifecycle phase, risk level, references, and approval status, ensuring thorough documentation and easy information retrieval via specific filtering. This structured approach facilitates organisational efficiency and enables the certification process to be adapted as required.

For practical use, the certification scheme we developed has been implemented as a base project; which can be readily exported, adapted, re-imported, or cloned to align with the particular requirements of the customer or AI system to be assessed. For certification bodies working with clients, the base project serves as the foundation from which the customer's certification project is derived. The customer's AIS is then assessed against the MOCs from the base scheme, supporting the issuance of the final certification.

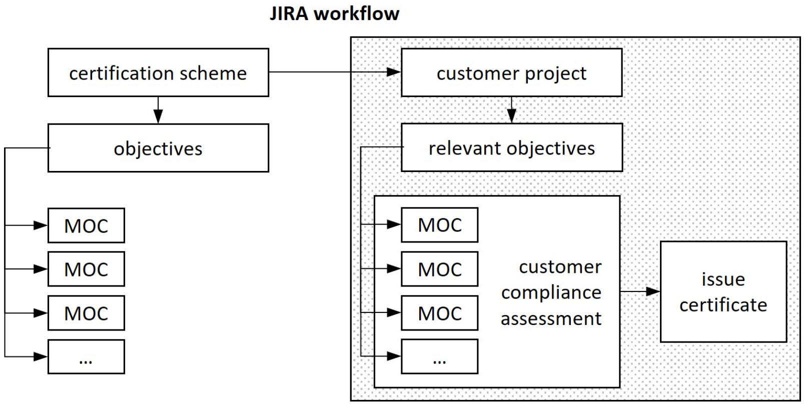

Figure 8 Visualisation of the certification workflow based on JIRA.

4.2 Machine Learning Operations Infrastructure

As argued in section 2.2. MLOps serves as enabler for trustworthy AI by design. It provides the necessary infrastructure and practices, ensuring that AIS are developed, deployed, and maintained reliably and efficiently. The adoption of MLOps thus facilitates the integration of trustworthy AI principles at every stage of the AI lifecycle, which are critical for regulatory compliance and societal acceptance [4]. MLOps extends DevOps practices to manage the complexities of bringing AIS into production, ensuring that they continuously meet trustworthiness standards [21].

The general architecture an AIS which adheres to MLOps principles supports the entire lifecycle of AIS and ensures that models are reproducible, reliable, and maintainable. This architecture includes project setup and requirements engineering, data engineering, model development, continuous integration/continuous deployment (CI/CD), and monitoring and maintenance. The requirement management system described in Section 4.1 will be part of the project setup and requirements engineering phase.

Complementing the requirements management, we implemented a full software pipeline for implementation, training and validation of AI models. This pipeline could be used (a) by AI developers for continuously tracking compliance with the certification requirements, or (b) by certifiers to perform systematic tests of their clients AI models. It thus demonstrates the benefits of MLOps best practices in terms of trustworthiness by design, implemented in the AIS development, as well as in terms of facilitating an efficient means of compliance tracking and verification as part of a certification.

In our pipeline, models are developed, trained and versioned using Git (through GitHub [13]) and MLflow [22], which document all changes to the models' code and parameters, respectively. MLflow also provides the tooling for tracking experiments, packaging code, and managing model deployment. The model development process involves experimentation with different algorithms and hyperparameters to optimise performance. GitHub Actions [14] and Apache Airflow [1] are used for workflow scheduling and monitoring and facilitate automated testing and deployment processes in CI/CD pipelines. Data is versioned using Oxen [30]. A schematic of the system is shown in Figure 9. The system listens for modifications to the model source code or input dataset. Changes automatically trigger training and model evaluation pipelines, which execute tests based on the methods described in sections 3.2.3, 3.2.4, and 3.3. For the certification scheme, we mainly relied on methods from captum, Alibi, AIX360, ART, and UQ360, as well as original implementations from academic papers. The

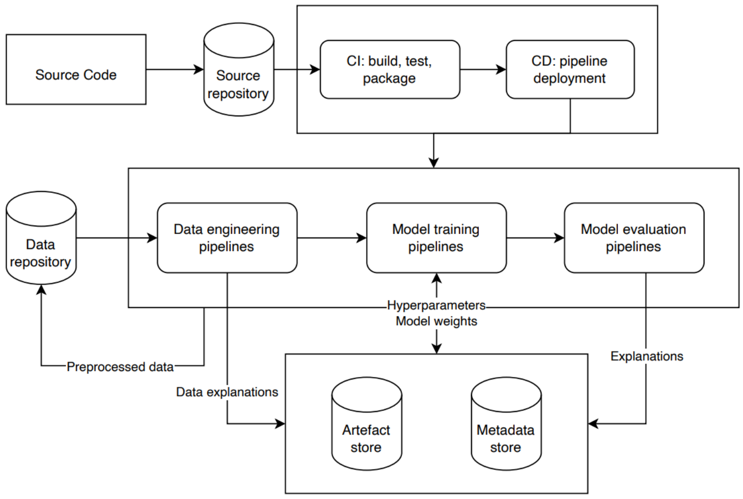

Figure 9 Overview of the MLOps system architecture.

outputs, model parameters and similar artefacts, are stored and versioned. Additionally, data engineering pipelines are run, which prepare the data for training and evaluation, and perform data-related trustworthiness evaluations.

MLOps provides several benefits to both AIS development and certification. Traceability and documentation are maintained throughout the AI lifecycle, providing a clear audit trail and ensuring that all objectives and means of compliance are systematically recorded. Version control is critical to maintaining the integrity of AI models and datasets, allowing teams to revert to previous versions if necessary and ensuring that all changes are documented and traceable. Automation and testing are streamlined through CI/CD pipelines, ensuring that each change is rigorously tested for compliance with trustworthiness standards before deployment. Post-deployment, continuous monitoring of AIS ensures that they remain compliant and perform reliably in real-world conditions.

Our workflow and methods have been tested in two real-world computer vision use cases in medical applications and vehicle detection on construction sites [10]. These use cases correspond to distinct high-risk applications according to the EU AI act. These use cases provide a test bed for validating the tools for certification on different data types and sets of requirements.

5 Discussion

The proposed certification scheme introduces several significant innovations in the assessment and certification of AIS trustworthiness, addressing an important gap in current practices. Despite the existence of standards, ethical guidelines and regulations, there remains a significant gap in the availability of practical tools and methodologies to achieve and systematically assess compliance. Our certification scheme addresses this gap by providing structured tools that are crucial for the rigorous evaluation of AIS. The scheme is underpinned by an extensive review and integration of 38 key documents from various standards and regulatory bodies, such as ISO/IEC, IEEE, EASA, and the Fraunhofer Institute. This foundational research ensures that the certification objectives and their means of compliance are comprehensive and aligned with the best practices and requirements across industries.

An important aspect of the scheme's implementation was evaluating multiple requirements management tools to support the certification workflow. Jira was selected for its robust capabilities in managing the complex certification process. This choice was crucial for maintaining systematic tracking of compliance objectives, ensuring that every requirement is meticulously documented and traceable.

Moreover, the means of compliance entail the application of metrics and technical methods by the customer, which can also be employed in the technical assessment of the AIS. Consequently, the scheme incorporates a technical assessment based on the implementation of selected technical methods which are linked to the objectives. The selection is determined through an evaluation of 95 well-established and cutting-edge methods, with the evaluation criteria being their suitability in meeting the defined objectives, criteria, and metrics. These methods were rigorously selected and empirically tested to ensure they provide effective compliance across various key aspects of trustworthiness, such as human oversight, transparency, safety, and (cyber)security. The workflow and methods developed within the certification scheme were tested on two real-life use cases: skin lesion classification and vehicle detection on construction sites. These practical applications demonstrate the scheme's effectiveness and adaptability in diverse, real-world scenarios. In addition, an automated workflow was implemented on a computing cluster following MLOps principles and best practices. This workflow maps MLOps stages with Trustworthy AI principles and key aspects, ensuring continuous compliance and efficient lifecycle management. By automating the certification process, the scheme enhances reliability, reduces human error, and ensures that the certification remains up-to-date with the latest developments in AI and ML technologies. Also, due to the dynamic nature of AIS and their complex post-deployment environments, trust levels can fluctuate. Continuous risk monitoring is essential to maintain trustworthiness, which is in line with the iterative nature of MLOps and is driven by versioning, automation, testing, deployment, and monitoring. Incorporating trustworthiness metrics alongside traditional performance metrics enables continuous feedback loops that systematically address trustworthiness requirements throughout the AI lifecycle [48].

The primary focus at the beginning of the development of the certification scheme was on reliability and transparency, areas where technical implementations could be more straightforwardly automated. As the scheme has developed, the scope has been expanded to encompass additional key areas, such as human oversight, which present more intricate challenges. These aspects are inherently linked to human interaction, which makes them challenging to automate effectively. The absence of established technical methods and metrics in these areas presents a significant challenge. As an illustration, the assessment of fairness in AI systems is an evolving field with no universally accepted metrics. This makes the certification process more challenging. The scheme provides a structured approach to compliance, whether through design or iterative testing and improvement. However, the absence of reliable metrics makes the implementation process less clear.

The tools and frameworks employed in the implementation of the certification scheme are designed to be adaptable, allowing the scheme to evolve in response to advances in AI techniques and changing requirements. One clear example is the increasing adoption of foundational models (referred to as general-purpose AI models in the EU legislation), including large language models (LLMs). These models, which are trained on vast and diverse datasets, introduce significant complexity due to their context-dependent and sometimes unpredictable behaviour. The subjective nature of their outputs and the difficulty of quantifying their decision-making processes pose challenges for evaluating and validating their trustworthiness within a standardised framework. As these models are increasingly deployed across many use cases, the development of new requirements, MOCs, and methods tailored to these models will be vital. Addressing these challenges will be essential for maintaining the relevance and applicability of the certification scheme as AI technologies continue to advance rapidly.

References

1 Apache Software Foundation. Airflow. URL: `https://airflow.apache.org/`.

2 Atlassian. Jira, 2002. URL: `https://www.atlassian.com/software/jira`.

3 Atlassian. Confluence, 2004. URL: `https://www.atlassian.com/software/confluence`.

4 Yann Billeter, Philipp Denzel, Ricardo Chavarriaga, Oliver Forster, Frank-Peter Schilling, Stefan Brunner, Carmen Frischknecht-Gruber, Monika Ulrike Reif, and Joanna Weng. MLOps as enabler of trustworthy AI. In *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*, 2024. `doi:10.21256/zhaw-30443`.

5 Jace Browning and Robert Adams. Doorstop: Text-based requirements management using version control, 2014. `doi:10.4236/jsea.2014.73020`.

6 Captum. Model interpretability for pytorch, 2023. URL: `https://captum.ai/`.

7 Chun Sik Chan, Huanqi Kong, and Guanqing Liang. A comparative study of faithfulness metrics for model interpretability methods. *arXiv preprint arXiv:2204.05514*, 2022.

8 Council of European Union. Laying down harmonized rules on artificial intelligence com(2021)206 final, 2021. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`.

9 Council of European Union. Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for AI in the World, 2023. URL: `https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/`.

10 Philipp Denzel, Stefan Brunner, Yann Billeter, Oliver Forster, Carmen Frischknecht-Gruber, Monika Ulrike Reif, Frank-Peter Schilling, Joanna Weng, Ricardo Chavarriaga, Amin Amini, et al. Towards the certification of ai-based systems. In *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*, 2024. `doi:10.21256/zhaw-30439`.

11 DIN, DKE. Artificial intelligence standardization roadmap, 2023. URL: `https://www.dke.de/en/areas-of-work/core-safety/standardization-roadmap-ai`.

12 EASA and Daedalean. Concepts of Design Assurance for Neural Networks (CoDANN) II. Technical report, May 2021. URL: `https://www.easa.europa.eu/en/downloads/128161/en`.

13 GitHub, Inc. GitHub . URL: `https://github.com/`.

14 GitHub, Inc. GitHub Actions. URL: `https://github.com/features/actions`.

15 Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online, June 2021. Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2021.naacl-demos.6`.

16 Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL: `http://jmlr.org/papers/v24/22-0142.html`.

17 IEEE. IEEE CertifAIEd: the mark of AI ethics, 2022. URL: `https://engagestandards.ieee.org/ieeecertifaied.html`.

18 IEEE Standards Association. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2023. URL: `https://standards.ieee.org/industry-connections/ec/autonomous-systems/`.

19 ISO. ISO/IEC JTC 1/SC 42 Artificial Intelligence, 2023. URL: `https://www.iso.org/committee/6794475.html`.

20 Anna Jobin, Marcello Ienca, and Effy Vayena. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. doi:10.1038/s42256-019-0088-2.

21 Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*, 11:31866–31879, 2023. doi:10.1109/ACCESS.2023.3262138.

22 LF Projects, LLC. MLFlow. URL: https://mlflow.org/.

23 LNE. Certification of processes for AI, 2023. URL: https://www.lne.fr/en/service/certification/certification-processes-ai.

24 Beatriz M. A. Matsui and Denise H. Goya. Mlops: A guide to its adoption in the context of responsible ai. In *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, pages 45–49, 2022. doi:10.1145/3526073.3527591.

25 Microsoft. InterpretML. URL: https://github.com/interpretml/interpret.

26 Microsoft. Azure devops, 2005. URL: https://azure.microsoft.com/en-us/products/devops/#overview.

27 NIST. NIST Technical AI Standards, 2023. URL: https://www.nist.gov/artificial-intelligence/technical-ai-standards.

28 NIST. AI Risk Management Framework (AI RMF) Knowledge Base, 2024. URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.

29 Independent High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. Technical report, European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

30 Oxen.ai. oxen. URL: https://www.oxen.ai/.

31 Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

32 Maximilian Poretschkin et al. Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog), 2021. URL: https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html.

33 Rational Software. Ibm doors, 2018. URL: https://www.ibm.com/docs/en/engineering-lifecycle-management-suite/doors.

34 Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. doi:10.21105/joss.02607.

35 SAE. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE J3016, 2021. URL: https://www.sae.org/standards/content/j3016_202104/.

36 Wojciech Samek et al. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. doi:10.1007/978-3-030-28954-6.

37 Anna Schmitz, Michael Mock, Rebekka Görge, Armin B Cremers, and Maximilian Poretschkin. A global scale comparison of risk aggregation in ai assessment frameworks. *AI and Ethics*, pages 1–26, 2024.

38 Seldon. Alibi Explain. URL: https://github.com/SeldonIO/alibi.

39 sicara. TF-Explain: Interpretability Methods for tf.keras Models with Tensorflow 2.x. URL: https://github.com/sicara/tf-explain.

40 Siemens. Polarion, 2004. URL: https://polarion.plm.automation.siemens.com/.

41 Ryan Soklaski, Justin Goodwin, Olivia Brown, Michael Yee, and Jason Matterer. Tools and practices for responsible ai engineering. *arXiv preprint arXiv:2201.05647*, 2022.

42 Guillaume Soudain. First usable guidance for Level 1 machine learning applications: A deliverable of the EASA AI Roadmap, 2021. URL: https://www.easa.europa.eu/en/downloads/134357/en.

43 The White House. Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, 2023. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

44 Trusted-AI LF AI Foundation. Adversarial Robustness Toolbox (ART). URL: https://github.com/Trusted-AI/adversarial-robustness-toolbox.

45 Trusted-AI LF AI Foundation. AI Explainability 360 (AIX360). URL: https://github.com/Trusted-AI/AIX360.

46 Trusted-AI LF AI Foundation. AI Uncertainty Quantification 360 (UQ360). URL: https://github.com/Trusted-AI/UQ360.

47 Arnaud Van Looveren et al. Alibi detect: Algorithms for outlier, adversarial and drift detection, 2019. URL: https://github.com/SeldonIO/alibi-detect.

48 Larysa Visengeriyeva, Anja Kammer, Isabel Bär, Alexander Kniesz, and Michael Plöd. MLOps Principles, 2020. URL: https://ml-ops.org/content/mlops-principles.