# Towards the Assessment of Trustworthy AI: A Catalog-Based Approach

Marco Anisetti[1], Claudio A. Ardagna[1], Nicola Bena[1] and Aneela Nasim[1,*]

[1]*Department of Computer Science, Università degli Studi di Milano, Milan, Italy*

**Abstract**

Artificial Intelligence (AI)-based systems are experiencing widespread adoption across a broad range of applications, including critical domains such as law and healthcare. This paradigm shift prompted a push towards the development of trustworthy AI systems, which are increasingly mandated by law and regulations. However, assessment techniques that concretely verify the trustworthiness of AI-based systems are still lacking. Current techniques in fact focus on traditional quality properties, providing either high-level guidelines or low-level techniques that cannot be generalized, and are therefore not applicable to AI-based systems. In this paper, we propose an assessment scheme that builds on a structured catalog of non-functional properties. The support for specific non-functional properties is verified along the entire system life cycle, from data collection to evaluation, by a set of assessment controls.

**Keywords**

Artificial Intelligence, Assessment, Non-functional property, Trustworthy AI

## 1. Introduction

Artificial Intelligence (AI) is gaining momentum, showcasing its growing importance across industries. The World Economic Forum states that AI will produce 97 million new jobs by 2025 and is expected to contribute trillions of dollars to the global economy by 2030 [1] highlighting its importance. AI is used to increase the efficiency and quality of processes in a plethora of tasks and domains, from industry 5.0 [2], to cybersecurity [3], health [4, 5], legal [6], and even military operations [7], to name but a few. These advancements are pointing towards *AI-based systems* (AI systems in the following), that is, distributed systems where AI models are used to implement end-user functionalities and manage system life cycle [8].

At the same time, awareness is mounting over the need for trustworthy AI systems, in terms of fairness, reliability, transparency, robustness, and privacy [9]. This demand is amplified in safety-critical domains, where mistakes and uncertainties in AI responses could have significant consequences [10]. However, claiming trustworthiness without evidence may negatively impact the system validity and user trust. Scholars (e.g., [11]) have indeed highlighted the need to *assess* the trustworthiness of AI systems, and assessment schemes are becoming essential and mandated by law (e.g., EU AI Act [12]).

State-of-the-art assessment schemes are largely inadequate to address this urgent need. On the one hand, AI assessment schemes (e.g., [13]) typically focus on functional properties, overlooking essential non-functional properties such as privacy, fairness, robustness. On the other hand, they are hardly generalizable, focusing only on a subset of the AI life cycle (i.e., dataset quality [14]) and specific properties (e.g., privacy, fairness) and domains (e.g., healthcare, law).

*TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.*

*Corresponding author.

marco.anisetti@unimi.it (M. Anisetti); claudio.ardagna@unimi.it (C. A. Ardagna); nicola.bena@unimi.it (N. Bena); aneela.nasim@unimi.it (A. Nasim)

https://anisetti.di.unimi.it (M. Anisetti); https://ardagna.di.unimi.it (C. A. Ardagna); https://homes.di.unimi.it/bena (N. Bena)

0000-0002-5438-9467 (M. Anisetti); 0000-0001-7426-4795 (C. A. Ardagna); 0000-0003-4909-9892 (N. Bena); 0009-0007-8204-0442 (A. Nasim)

The assessment scheme in this paper aims to initially address the above gaps. Our scheme codifies best practices into a structured *catalog* of non-functional properties relevant for trustworthy AI (i.e., reliability, transparency, fairness, robustness, privacy). Each non-functional property is linked to a set of *controls* that span the entire AI system life cycle and verify whether the AI system supports the required non-functional properties. Our scheme integrates evidence collected in each phase of the life cycle to provide a complete assessment of system trustworthiness.

Our contribution is twofold. First, we introduce a catalog that systematically organizes well-known non-functional properties and related controls insisting on the whole the AI system life cycle, from data collection to AI inference (Section 4). Second, we implement an assessment process that selects the most suitable set of properties to be assessed and related control on the basis of the target system peculiarities, and collects and analyzes evidence accordingly, thereby supporting a reproducible assessment (Section 3).

## 2. Background and Motivations

Assessment schemes have been defined since the 80s to verify whether IT systems behave as expected and meet desired non-functional requirements (properties) [15]. Assessment schemes are based on an *assessment model* that defines the activities that have to be executed to prove that a target system supports a given *non-functional property* according to a set of *evidence* collected following the assessment model. If the evidence is successfully collected, a *compliance report* is issued, for instance in the form of a certificate [15]. Several techniques can be used for assessment, such as certification [16], stress testing [17], audits [18], to name but a few.

With the growing need for assessment schemes, researchers developed new schemes in parallel with the advancement of IT systems, initially targeting traditional software systems (e.g., [19]) and later extended towards cloud (e.g., [20]) and network (e.g., [21]) services, and IoT systems [22]. In addition, assessment schemes have been developed for software produced using waterfall and agile methodologies (e.g., [23]) in accordance to standards such as ISO/IE [24] and DO-178C [25].

As technology advanced, the scope of assessment schemes expanded beyond traditional IT systems to address the unique challenges posed by the complex nature of AI systems, that are defined as (distributed) IT systems where AI models play the key roles of providing end-user functionalities and managing the system life cycle [8]. Early research on the topic primarily highlighted the importance of end-to-end transparency [26] and reproducibility [27] of the entire AI life cycle. Research also focused on dataset quality, emphasizing the importance of robust dataset definition and validation methods to ensure that the data used to train the AI model are accurate, representative, and bias-free [28]. As AI models are being deployed in safety-critical and high-stakes environments, the attention shifted towards non-functional properties at the basis of trustworthy AI, often outside the context of a concrete assessment scheme. These non-functional properties include robustness, privacy, and security [29]. For instance, robustness has been extensively investigated (e.g., [30]), particularly in terms of protection against poisoning and evasion attacks (e.g., [31]). At the same time, concerns over biases prompted research on the assessment of fairness (e.g., [32]). To further support AI trustworthiness, life cycle management tools (e.g., *MLflow* [33]) have been adopted to ensure structured model documentation, reproducibility, traceability, and reliable and auditable deployment pipelines [34].

Despite the research advances on the assessment of AI systems, significant limitations remain. On the one hand, existing schemes are hardly applicable to AI systems in their entirety as they primarily target traditional software components, while AI-specific schemes focus solely on assessing AI models [35]. Furthermore, AI assessment schemes typically overlook non-functional properties, which are increasingly mandated by law (e.g., EU AI Act) [36]. This results in inefficiencies, legal uncertainties, non-compliance, and bias. In addition, existing AI assessment schemes mainly focus on dataset quality, neglecting other crucial phases such as training and evaluation. The exclusion of critical aspects, such as overfitting [37] at the training phase and inappropriate performance measures [38] at the evaluation phase, can lead to false positives during system assessment. Finally, existing AI assessment schemes

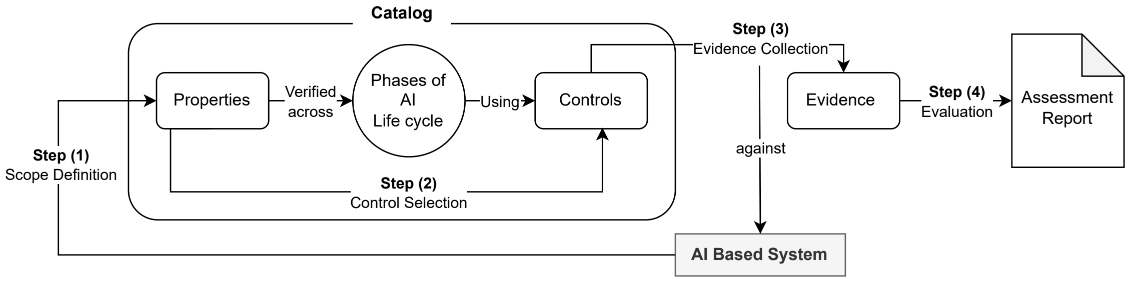


**Figure 1:** Overview of our catalog-based assessment scheme.

are hardly generalizable. They are defined for specific domains or properties (e.g., healthcare [4], fairness [32]). This fragmentation limits the applicability and scalability of AI assessment schemes across different sectors.

The scheme in this paper provides a first solution to these issues, addressing the need for a unified and generalizable scheme for the assessment of non-functional properties of AI systems.

## 3. Our Approach

Figure 1 illustrates an overview of our assessment scheme for the non-functional assessment of AI systems. It is built on a *catalog* of *non-functional properties* and associated *controls* (Section 4). The catalog defines a set of non-functional properties modeling the expected system behavior. The catalog also includes a set of controls; each control assesses specific aspects of the system that contribute to the support of one or more non-functional properties. The controls are applied at different phases of the AI system life cycle and collect *evidence* to verify the AI system in its entirety.

The assessment of an AI system according to our scheme consists of four steps.

- **Step (1): Scope definition.** It defines the assessment scope by selecting and configuring the relevant (set of) non-functional property from the catalog in Section 4. The selection depends on the system peculiarities, domain criticality (e.g., healthcare vs. retail), legal and regulatory requirements (e.g., the risk level according to the EU AI Act [12]), and the objectives of the system owner. For instance, in critical domains, Step (1) prioritizes reliability and fairness, whereas in consumer applications, it emphasizes transparency. Step (1) also fixes the appropriate interpretation of each selected property and configures it accordingly, since different properties often have different definitions depending on the context [11]. Step (1) finally clarifies when and where each property should be assessed during the AI life cycle. For instance, certain properties may need to be assessed during data collection (e.g., fairness), while others are more relevant at the AI model evaluation phase (e.g., transparency).

- **Step (2): Control selection.** It identifies the appropriate controls from the catalog in Section 4, starting from the defined scope. Each control assesses specific aspects of the AI system to check whether the selected (set of) non-functional property is supported. The catalog maps these controls according to the target properties and the phases of the AI life cycle (e.g., data collection, training, evaluation). Controls are then configured in alignment with the system characteristics (e.g., use case, property, life cycle) where they operate. Controls are then configured according to the scope identified in Step (1).

- **Step (3): Evidence collection.** It executes each control selected at Step (2) according to the assessment scope. Each control collects a set of *evidence*, which is the basis to assess the AI system. Evidence can take various forms depending on how controls have been configured and

**Table 1**
The catalog of our assessment scheme.

| Life Cycle Phase | Controls | Properties |
|---|---|---|
| Data Collection | Data Integrity | Fairness, Reliability, Robustness |
| | Data Exposure | Privacy, Robustness, Reliability, Fairness |
| | Balanced Dataset | Fairness, Reliability, Robustness |
| | Dataset Diversity | Reliability, Robustness, Fairness |
| | Sampling Bias | Fairness, Reliability, Robustness |
| | Data Uniqueness | Reliability, Robustness,Fairness |
| | Label Integrity | Reliability, Robustness, Fairness |
| | Handling Of Sensitive Data | Privacy |
| Training | Overfitting and Underfitting | Reliability, Robustness, Fairness |
| | Spurious Correlations | Reliability, Robustness, fairness |
| | Data Snooping | Reliability, Robustness, Fairness |
| | Parameter Selection | Reliability,Fairness, Robustness |
| | Explainability | Transparency |
| Evaluation | Appropriate Baseline | Reliability |
| | Appropriate Performance Measures | Reliability |
| | Base Rate Fallacy | Reliability, Robustness |
| | Transparency In Results | Transparency |
| | Performance Consistency | Reliability, Robustness |
| | Testing On Unprocessed Data | Reliability, Robustness |
| | Assessing confidence intervals and variance | Reliability |
| | Reproducible ML lifecycle | Transparency |

implemented. For instance, evidence can be system logs (e.g., training checkpoints), performance metrics (e.g., accuracy values), documentation, or direct observation of system behavior during its execution.

- **Step (4): Evaluation.** It analyzes each collected evidence against the control-specific criteria defined in the previous steps, and aggregates these outcomes across the different considered phases of the AI life cycle phases. Analysis determines the extent to which the AI system satisfies the non-functional properties selected at Step (1), according to the identified scope. Based on this evaluation, a positive or negative compliance decision is finally made. In case the decision is positive, a *compliance report* is issued, detailing the properties, controls, and corresponding evidence, thus supporting transparency and reproducibility. If the decision is negative, evidence can be used as source of remediation.

## 4. The Catalog

Our catalog includes five non-functional properties (i.e., reliability, transparency, fairness, privacy, and robustness) and corresponding controls for AI assessment. With no lack of generality, we focus on the most common definitions of the properties that consistently appear as core requirements for trustworthy AI in major guidelines and regulations, including the EU AI Act, the NIST AI Risk Management Framework, and recent surveys [9, 39, 40].

- **Reliability** refers to the AI system's ability to consistently perform as intended throughout its operational life cycle, ensuring stability and the capacity to withstand failures without significant loss in decision-making process [39]. For instance, data diversity should be prioritized to ensure that the model is trained on representative data across all classes.

- **Transparency** refers to the AI system ability to make its decision-making process understandable and explained to stakeholders (e.g., users, developers, or regulators) [41] so that the reasons behind the decisions taken by the system can be traced and explained.

- **Privacy** refers to the AI system ability to protect sensitive data from unauthorized access, misuse, and leakage, ensuring that personal information is securely handled throughout the system life cycle, from data collection to training and inference [42]. For instance, the presence of individual data points in the training set should not be inferred.

- **Fairness** refers to the AI system ability to avoid any favoritism in decision-making toward an individual or group based on their inherent or acquired characteristics (e.g., race, gender), ensuring careful data collection and training on diverse and representative datasets [40].

- **Robustness** refers to the AI system ability to maintain performance and accurate decision-making when exposed to variations and unexpected inputs, ensuring it can adapt to changes in the environment or input data without degradation. This ability can be compromised at various phases of the AI life cycle if not properly managed [43].

We note that the interpretation of these properties, fixed during Step (1) in Section 3, may vary significantly across domains. For instance, in the healthcare domain, fairness is often defined as demographic equity in diagnosis, and reliability as the ability to provide consistent results across diverse patient groups and imaging modalities (see Example 1). As another example, in finance, fairness is often defined as equal treatment in loan decisions [44]. Moreover, conflicts between properties may arise. For instance, stronger privacy can reduce transparency [45]. Such conflicts must be addressed according to the prioritized requirements of the application domain and applicable regulations.

The other component of our catalog is controls, linked to the three phases of the AI life cycle (data collection, training, and evaluation) on the basis of their relevance to assess the given properties. Controls can be implemented through automated checks (e.g., detecting overfitting using validation metrics), while others require manual review (e.g., inspecting label integrity in datasets), or statistical validation (e.g., checking for sampling bias).

Table 1 shows the mapping between the phases of the AI life cycle, the controls, and the properties assessed by those controls, which are detailed in Table 2. This mapping has been designed by focusing on the possible issues that may arise, and in turn invalidate, the required non-functional properties, during all the phases of the AI life cycle (e.g., [46]). For instance, fairness can be assessed through controls *Balanced dataset*, *Sampling Bias*, and *Label Integrity*, while robustness can be assessed through controls *Overfitting and Underfitting*, *Spurious Correlations*, and *Performance Consistency*.

**Example 1.** *We illustrate the application of our catalog-based assessment scheme to an AI system in the healthcare domain. The system is built on Federated Learning (FL) for segmenting tumor masses from MRI (Magnetic Resonance Imaging) images. Being healthcare a high-risk domain according to the EU AI Act [12] the system* will have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the Union market*". Among the mandatory requirements, reliability assumes critical importance to ensure that the system provides consistent and accurate results in a setting where data can vary significantly across patients, imaging modalities, and environmental conditions. To assess whether the FL-based system supports property reliability, we follow the assessment process in Section 3, first defining the assessment scope that covers all phases of the system life cycle, and then selecting controls accordingly.*

*The assessment in phase data collection focuses on verifying the adequacy and structure of the datasets. In this context, the selected control* Data Diversity *inspects the diversity of the datasets. It verifies whether the datasets contain sufficient samples, acquired using different MRI modalities and collected from the most widely used MRI equipments. The outcome of this control is successful if the collected evidence shows that at least three MRI vendors and* T2/DWI/ADC *modalities have been used.*

*The assessment in phase training focuses on verifying the level of generalization of the federated AI model. In this context, the selected control* Parameter Selection *inspects the training algorithm to verify whether hyperparameters have been systematically tuned without bias, avoiding over-optimization. The outcome of this control is successful if the collected evidence shows that the* Bayesian search strategy *has been used. This strategy implements a systematic approach that reduces the risk of biased performance outcomes.*

**Table 2**
Controls of our catalog.

| Name | Description |
|---|---|
| **Data Integrity** | Verifies the completeness and quality of dataset by ensuring no missing data entries or variables. |
| **Data Exposure** | Verifies data partitioning and integrity to avoid unintended data leakage from outside the training dataset. |
| **Balanced Dataset** | Verifies that the distribution of classes or categories within the dataset is balanced to avoid bias. |
| **Dataset Diversity** | Verifies that the dataset represents diverse demographics and scenarios to avoid performance deficiencies. |
| **Sampling Bias** | Verifies whether the collected data fairly represents the target population to avoid bias. |
| **Data Uniqueness** | Identifies redundant or repeated records in the dataset that may lead to biased learning or overfitting. |
| **Label Integrity** | Verifies the correctness of data labeling to avoid incorrect annotations. |
| **Handling Of Sensitive Data** | Verifies that sensitive personal information is properly handled and protected. |
| **Overfitting and Underfitting** | Verifies that the model has achieved generalization and learned the underlying patterns. |
| **Spurious Correlations** | Verifies that the model is learning accurate associations and not being influenced by irrelevant variables. |
| **Data Snooping** | Ensures that the training and model selection processes are free from unintended bias or data leakage from the test set. |
| **Parameter Selection** | Verifies that the tuning of model parameters is conducted without bias, avoiding over-optimization. |
| **Explainability** | Verifies that the model's decision-making process is interpretable and understandable. |
| **Appropriate Baseline** | Verifies that the model is evaluated against valid reference model to avoid overstated performance metrics. |
| **Appropriate Performance Measures** | Verifies that performance metrics align with the task's objectives, ensuring a meaningful evaluation. |
| **Base Rate Fallacy** | Verifies that the statistical prevalence of events is properly considered to avoid misleading conclusions. |
| **Transparency in Results** | Verifies the clarity and traceability of model outputs to ensure they can be understood and audited. |
| **Performance Consistency** | Verifies that the model's performance is consistent across different datasets and environments. |
| **Testing on Unprocessed Data** | Verifies that test data includes real-world imperfections, ensuring a realistic evaluation. |
| **Assessing Confidence Intervals and Variance** | Verifies that confidence intervals and variance are considered during model evaluation. |
| **Reproducible ML Lifecycle** | Verifies that the ML lifecycle is replicable, ensuring the model's results are consistent and audit-able. |

*The assessment in phase evaluation focuses on verifying the performance of the AI model in real-world conditions. In this context, the selected controls* Appropriate Baseline *and* Appropriate Performance Measure *inspect the evaluation process to verify whether the federated AI model has been compared against a centralized model built with the same model architecture and using an adequate evaluation metric. The outcome of this control is successful if the collected evidence shows that a comparison process has been executed.*

# 5. Conclusions

As AI-based systems are increasingly integrated into critical applications, the need to gain confidence in their behavior becomes fundamental. Existing assessment schemes, however, are largely inadequate to address this challenge, thereby undermining system trustworthiness and, consequently, end-user acceptance. In this paper, we proposed a preliminary scheme for the non-functional assessment of AI systems. The scheme is based on a catalog that binds properties and controls to assess the target AI system along its entire life cycle.

Future work includes expanding the catalog to include additional properties and controls, enabling a more comprehensive assessment of AI-based system behaviors. Additionally, we will focus on composite AI-based systems, where multiple and diverse (AI-based) services are jointly used to implement the system functionalities and manage its life cycle. In this case, our catalog can be extended to assess individual AI services, and the retrieved results combined with those retrieved using traditional assessment schemes. A final aggregation step can then jointly analyze the collected evidence and produce an overall compliance report. Finally, we will pursue practical evaluations on real word AI based systems to validate the scheme's effectiveness in operational settings.

## Acknowledgments

Research supported, in parts, by *i)* project BA-PHERD, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.1: "Fondo Bando PRIN 2022" (CUP G53D23002910006); *ii)* MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D" (CUP G43C22001370007, Code ECS00000037); *iii)* project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NextGenerationEU. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor the Italian MUR can be held responsible for them.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

[1] World Economic Forum, The Future of Jobs Report 2020, Technical Report, World Economic Forum, 2020. URL: https://www.weforum.org/publications/the-future-of-jobs-report-2020/.

[2] L. Colombi, M. Vespa, N. Belletti, M. Brina, S. Dahdal, F. Tabanelli, F. Resca, E. Bellodi, M. Tortonesi, C. Stefanelli, M. Vignoli, Embedding Models for Multivariate Time Series Anomaly Detection in Industry 5.0, Data Science and Engineering (2025).

[3] M. Atzori, E. Calò, L. Caruccio, S. Cirillo, G. Polese, G. Solimando, Evaluating password strength based on information spread on social networks: A combined approach relying on data reconstruction and generative models, Online Social Networks and Media 42 (2024).

[4] S. Cirillo, S. C. Rajkumar, L. Solimando, D. Yuvasini, A hybrid approach combining images and questionnaires for early detection and severity assessment of Autism Spectrum Disorder, Image and Vision Computing 160 (2025).

[5] V. Bevilacqua, A. Di Marino, E. Di Nardo, A. Ciaramella, I. De Falco, G. Sannino, Cross-domain Super-Resolution in Medical Imaging, in: Proc. of IEEE ISCC 2024, Paris, France, 2024.

[6] V. Bellandi, S. Maghool, S. Siccardi, An NLP-based statistical reporting methodology applied to court decisions, in: Proc. of Euromicro SEAA 2023, Durres, Albania, 2023.

[7] L. Colombi, S. Dahdal, E. D. Caro, R. Fronteddu, A. Gilli, Efficient Data Dissemination via Semantic Filtering at the Tactical Edge, in: Proc. of IEEE MILCOM 2024, Washington, DC, USA, 2024.

[8] J. Bogner, X. Franch, S. Martínez-Fernández, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, S. Wagner, Software engineering for AI-based systems: a survey, ACM Transactions on Software Engineering and Methodology 31 (2022).

[9] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM Computing Surveys 55 (2022).

[10] F. T. S. Chan, E. L. Droguett, T. Han, A. Mosleh, L. Zhang, T. Zhou, An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications, Reliability Engineering & System Safety 229 (2023).

[11] M. Anisetti, C. A. Ardagna, N. Bena, E. Damiani, Rethinking certification for trustworthy machine-learning-based applications, IEEE Internet Computing 27 (2023).

[12] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM Document COM/2021/206 final, European Commission, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[13] M. Al-Attar, A. R. Brentnall, J. Cuzick, C. Damiani, G. Kalliatakis, E. F. Lane, G. Montana, C. Pudney, J. Rose, M. Sreenivas, Evaluation of an AI model to assess future breast cancer risk, Radiology 307 (2023).

[14] Y. Gong, R. Li, G. Liu, L. Meng, Y. Xue, A survey on dataset quality in machine learning, Information and Software Technology 162 (2023).

[15] C. A. Ardagna, N. Bena, Non-Functional Certification of Modern Distributed Systems: A Research Manifesto, in: Proc. of IEEE SSE 2023, Chicago, IL, USA, 2023.

[16] E. Ilyushin, D. Namiot, On Certification of Artificial Intelligence Systems, Physics of Particles and Nuclei 55 (2024).

[17] J. Li, M. McCallen, B. Moeini, S. Nejati, M. Sabetzadeh, A Lean Simulation Framework for Stress Testing IoT Cloud Systems, IEEE Transactions on Software Engineering 50 (2024).

[18] T. Behrend, R. N. Landers, Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models., American Psychologist 78 (2023).

[19] C. Baron, V. Louis, Framework and tooling proposals for Agile certification of safety-critical embedded software in avionic systems, Computers in Industry 148 (2023).

[20] A. Martin, Y. Nugraha, Towards a framework for trustworthy data security level agreement in cloud procurement, Computers & Security 106 (2021).

[21] M. Anisetti, C. A. Ardagna, F. Berto, E. Damiani, A security certification scheme for information-centric networks, IEEE Transactions on Network and Service Management 19 (2022).

[22] M. Anisetti, C. A. Ardagna, N. Bena, R. Bondaruc, Towards an Assurance Framework for Edge and IoT Systems, in: Proc. of IEEE EDGE 2021, Guangzhou, China, 2021.

[23] A. Aguiar, J. Ribeiro, J. G. Silva, Beyond tradition: evaluating agile feasibility in DO-178C for aerospace software development, arXiv preprint arXiv:2311.04344 (2023).

[24] International Organization for Standardization, ISO/IEC 25002:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model, Technical Report, International Organization for Standardization, 2014. URL: https://www.iso.org/standard/35746.html.

[25] D. Dollinger, K. Dmitriev, M. Hochstrasser, F. Holzapfel, Y. Lai, S. Myschik, P. Nagarajan, M. Saleab, K. Schmiechen, S. A. Zafar, A lean and highly-automated model-based software development process based on do-178c/do-331, in: Proc. of IEEE DASC 2020, San Antonio, TX, USA (held virtually), 2020.

[26] T. Toy, Transparency in AI, AI & SOCIETY 39 (2024).

[27] H. J. W. L. Aerts, A. Barberis, F. M. Buffa, Robustness and reproducibility for AI learning in biomedical sciences: RENOIR, Scientific Reports 14 (2024).

[28] I. Caballero, F. Gualo, M. Piattini, M. Rodríguez, J. Verdugo, Data quality certification using ISO/IEC 25012: Industrial experiences, Journal of Systems and Software 176 (2021).

[29] D. Elliott, E. Soifer, AI technologies, privacy, and security, Frontiers in Artificial Intelligence 5 (2022).

[30] A. Balayn, M. Brambilla, L. Corti, P. Lippmann, A. Tocchetti, J. Yang, M. Yurrita, Ai robustness: a human-centered perspective on technological challenges and opportunities, ACM Computing Surveys 57 (2025).

[31] M. Anisetti, C. A. Ardagna, N. Bena, E. Damiani, C. Y. Yeun, Protecting machine learning from poisoning attacks: A risk-based approach, Computers & Security 155 (2025).

[32] R. J. Chen, T. Y. Chen, J. Lipkova, M. Y. Lu, F. Mahmood, S. Sahai, J. J. Wang, D. F. K. Williamson, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nature biomedical engineering 7 (2023).

[33] MLflow Project, MLflow: An Open Source Platform for the Machine Learning Lifecycle, 2025. URL: https://mlflow.org/.

[34] V. Vangala, MLOps in Practice: A Framework for Scalable AI Model Deployment, Monitoring, and Retraining, International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence 13 (2022).

[35] M. Everett, B. Lütjens, Certifiable robustness to adversarial state uncertainty in deep reinforcement

learning, IEEE Transactions on Neural Networks and Learning Systems 33 (2021).
[36] C. A. Ardagna, M. Anisetti, N. Bena, G. Gianini, Certifying Accuracy, Privacy, and Robustness of ML-Based Malware Detection, SN Computer Science 5 (2024).
[37] S. Kaltenbrunner, C. Korab, P. H. Luz de Araujo, B. Roth, Y. Xia, Specification overfitting in artificial intelligence, Artificial Intelligence Review 58 (2025).
[38] D. Arp, L. Cavallaro, F. Pendlebury, F. Pierazzi, E. Quiring, K. Rieck, A. Warnecke, C. Wressnegger, Pitfalls in Machine Learning for Computer Security, Communications of the ACM 67 (2024).
[39] S. T. H. Mortaji, M. E. Sadeghi, Assessing the reliability of artificial intelligence systems: Challenges, metrics, and future directions, International Journal of Innovation in Management, Economics and Social Sciences 4 (2024).
[40] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys 54 (2021).
[41] L. Montgomery, K. Kent, M. S. John, J. Greer, A. Stavrou, A. Joshi, A. Ray, R. Chandramouli, T. Oates, P. S. Bishnoi, J. A. Hogan, M. L. Badger, K. Kent, D. A. Boyd, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST Special Publication 1270, National Institute of Standards and Technology (NIST), 2023. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.
[42] J. García-Ortiz, W. Villegas-Ch, Toward a comprehensive framework for ensuring security and privacy in artificial intelligence, Electronics 12 (2023).
[43] I. Sutskever, C. Szegedy, W. Zaremba, et al., Intriguing properties of neural networks, arXiv:1312.6199 (2014).
[44] S. Raziyeva, M. Meraliyev, Bias and Fairness in Automated Loan Approvals: A Systematic Review of Machine Learning Approaches, Journal of Emerging Technologies and Computing 1 (2025).
[45] C. Sanderson, D. Douglas, Q. Lu, Implementing responsible AI: Tensions and trade-offs between ethics aspects, in: Proc. of IEEE IJCNN 2023, Gold Coast, Australia, 2023.
[46] D. Arp, L. Cavallaro, F. Pendlebury, F. Pierazzi, E. Quiring, K. Rieck, A. Warnecke, C. Wressnegger, Dos and don’ts of machine learning in computer security, in: Proc. of USENIX Security 2022, Boston, MA, USA, 2022.