

Symposium on Scaling AI Assessments

SAIA 2024, September 30–October 1, 2024, Cologne, Germany

Edited by

Rebekka Görge

Elena Haedecke

Maximilian Poretschkin

Anna Schmitz



Editors

Rebekka Göрге

Fraunhofer IAIS, Sankt Augustin, Germany
rebekka.goerge@iais.fraunhofer.de

Elena Haedecke

Fraunhofer IAIS, Sankt Augustin, Germany
University of Bonn, Bonn, Germany
elena.haedecke@iais.fraunhofer.de

Maximilian Poretschkin

Fraunhofer IAIS, Sankt Augustin, Germany
University of Bonn, Bonn, Germany
Maximilian.Poretschkin@iais.fraunhofer.de

Anna Schmitz

Fraunhofer IAIS, Sankt Augustin, Germany
Anna.Schmitz@iais.fraunhofer.de

ACM Classification 2012

Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning;
Applied computing; Social and professional topics → Computing / technology policy; General and
reference → General conference proceedings; General and reference → Cross-computing tools and
techniques; Software and its engineering → Software creation and management

ISBN 978-3-95977-357-7

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern,
Germany. Online available at <https://www.dagstuhl.de/dagpub/978-3-95977-357-7>.

Publication date

January, 2025

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed
bibliographic data are available in the Internet at <https://portal.dnb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0):
<https://creativecommons.org/licenses/by/4.0/legalcode>.



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work
under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/OASlcs.SAIA.2024.0

ISBN 978-3-95977-357-7

ISSN 1868-8969

<https://www.dagstuhl.de/oasics>

OASlcs – OpenAccess Series in Informatics

OASlcs is a series of high-quality conference proceedings across all fields in informatics. OASlcs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

Editorial Board

- Daniel Cremers (TU München, Germany)
- Barbara Hammer (Universität Bielefeld, Germany)
- Marc Langheinrich (Università della Svizzera Italiana – Lugano, Switzerland)
- Dorothea Wagner (*Editor-in-Chief*, Karlsruher Institut für Technologie, Germany)

ISSN 1868-8969

<https://www.dagstuhl.de/oasics>

■ Contents

Preface	
<i>Rebekka Göрге, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz</i>	0:vii
Organizers of the Workshop	
.....	0:xi

Safeguarding and Assessment Methods

On Assessing ML Model Robustness: A Methodological Framework	
<i>Afef Awadid and Boris Robert</i>	1:1–1:10
Trustworthy Generative AI for Financial Services	
<i>Marc-André Zöller, Anastasiia Iurshina, and Ines Röder</i>	2:1–2:5

Risk Assessment and Evaluations

EAM Diagrams – A Framework to Systematically Describe AI Systems for Effective AI Risk Assessment	
<i>Ronald Schnitzer, Andreas Hapfelmeier, and Sonja Zillner</i>	3:1–3:16
Scaling of End-To-End Governance Risk Assessments for AI Systems	
<i>Daniel Weimer, Andreas Gensch, and Kilian Koller</i>	4:1–4:5
Risk Analysis Technique for the Evaluation of AI Technologies with Respect to Directly and Indirectly Affected Entities	
<i>Joachim Iden, Felix Zwarg, and Bouthaina Abdou</i>	5:1–5:6
SafeAI-Kit: A Software Toolbox to Evaluate AI Systems with a Focus on Uncertainty Quantification	
<i>Dominik Eisl, Bastian Bernhardt, Lukas Höhndorf, and Rafal Kulaga</i>	6:1–6:3

Ethics and Standards

Towards Trusted AI: A Blueprint for Ethics Assessment in Practice	
<i>Christoph Tobias Wirth, Mihai Maftei, Rosa Esther Martín-Peña, and Iris Merget</i>	7:1–7:19
AI Readiness of Standards: Bridging Traditional Norms with Modern Technologies	
<i>Adrian Seeliger</i>	8:1–8:6

Governance and Regulations

Introducing an AI Governance Framework in Financial Organizations. Best Practices in Implementing the EU AI Act	
<i>Sergio Genovesi</i>	9:1–9:7

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Göрге, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz

OpenAccess Series in Informatics



ASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Evaluating Dimensions of AI Transparency: A Comparative Study of Standards, Guidelines, and the EU AI Act <i>Sergio Genovesi, Martin Haimerl, Iris Merget, Samantha Morgaine Prange, Otto Obert, Susanna Wolf, and Jens Ziehn</i>	10:1–10:17
--	------------

Transparency and XAI

Transparency of AI Systems <i>Oliver Müller, Veronika Lazar, and Matthias Heck</i>	11:1–11:7
A View on Vulnerabilities: The Security Challenges of XAI <i>Elisabeth Pacht, Fabian Langer, Thora Markert, and Jeanette Miriam Lorenz</i>	12:1–12:23

Certification

AI Certification: Empirical Investigations into Possible Cul-De-Sacs and Ways Forward <i>Benjamin Fresz, Danilo Brajovic, and Marco F. Huber</i>	13:1–13:4
AI Certification: An Accreditation Perspective <i>Susanne Kuch and Raoul Kirmes</i>	14:1–14:7
AI Assessment in Practice: Implementing a Certification Scheme for AI Trustworthiness <i>Carmen Frischknecht-Gruber, Philipp Denzel, Monika Reif, Yann Billeter, Stefan Brunner, Oliver Forster, Frank-Peter Schilling, Joanna Weng, and Ricardo Chavarriaga</i>	15:1–15:18

■ Preface

This volume presents scientific and practical contributions from the Symposium on Scaling AI Assessments (SAIA 2024). SAIA 2024 was held on September 30 and October 1, 2024 in Cologne, Germany. It gathered practitioners from the TIC sector (testing, inspection, certification), representatives from tech start-ups and AI deployers, as well as researchers in the field of trustworthy AI. Together, they discussed and promoted solution approaches towards scalable AI assessments.

Especially against the background of European AI regulation, AI conformity assessment procedures are of particular importance, both for specific use cases and for general-purpose models. But also in non-regulated domains, the quality of AI systems is a decisive factor as unintended behavior can lead to serious financial and reputation damage. As a result, there is a great need for AI audits and assessments and in fact, it can also be observed that a corresponding market is forming. At the same time, there are still (technical) challenges in conducting the required assessments and a lack of extensive practical experience in evaluating different AI systems. Overall, the emergence of the first marketable, commercial AI assessment offerings is just in the process and a definitive, distinct procedure for AI quality assurance has not yet been established. These outstanding challenges can be addressed from two perspectives which must be intertwined to enable scalable solutions:

- **Operationalization perspective:** AI assessments require further operationalization both at level of governance and related processes and at the product level. Empirical research is pending that applies and evaluates governance frameworks, assessment criteria, AI quality KPIs and methodologies in practice for different AI use cases.
- **Testing tools and implementation perspective:** Conducting AI assessments in practice requires a testing ecosystem and tool support, as many quality KPIs cannot be calculated without tool support. At the same time automation of such assessments is a prerequisite to make the corresponding business model scale.

Taking a pragmatic and market-oriented approach in bringing together the two perspectives, SAIA 2024 includes practitioner contributions in addition to academic papers. Specifically, the practitioner track was open for short abstracts of practice reports and case studies, some of which were extended to full papers after the conference. Regarding the academic track, SAIA 2024 places particular emphasis on the commitment of young researchers along more experienced participants. The detailed list of the topics of interest is provided below. Beyond the presentations from the academic and practitioner tracks, the conference program included keynotes by Prof. Dr. Bertrand Braunschweig, scientific coordinator of Confiance.ai, and Prof. Dr. Roberto V. Zicari, head of the Z-Inspection initiative, who shared their experience on implementing trustworthy and ethical AI in practice. In addition, a legal panel with Dr. Andreas Engel, Prof. Dr. Dimitrios Linardatos and Prof. Dr. Mark Cole dealt with questions such as what requirements the AI Act places on generative AI and how it interacts with other complementary legal frameworks such as the GDPR.

We thank the program committee very much for their contribution to the planning and organization of the Symposium on Scaling AI Assessments and for their effort in reviewing the papers with care and quality. We are especially grateful for the international cooperation in the program committee with representatives of Confiance.ai, Confiance IA and CSIRO Australia. With your support, SAIA 2024 provided a framework for practitioners and researchers the field of AI assessment to become more connected as an interdiscip-

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

linary community. In this sense, SAIA 2024 should also be seen as a contribution to the development of an international and interdisciplinary community on this topic, building on previous conferences and workshops, namely *AITA: AI Trustworthiness Assessment* and *RAIE: International Workshop on Responsible AI Engineering*^{1,2}. We hope that our joint efforts have encouraged further cooperation also beyond the conference, since this is an important prerequisite for driving scalable AI assessment forward and expanding the scientific state of art at the same time. Last but not least, we thank all the participants for presenting their work and contributing to lively discussions.

SAIA 2024 and these proceedings were organized as part of the flagship project ZERTIFIZIERTE KI which is funded by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia, Germany. The editors would like to thank the consortium for the successful cooperation.

Topics of Interest

- **Standardization of concepts and frameworks for AI assessment**
 - *Operationalization perspective:* How can basic concepts of AI assessments such as the target-of-evaluation, the operational environment and the operational design domain (ODD) be specified in a standardized way? How can compatibility with existing assessment or certification frameworks for other domains (e.g. safety, data protection) be guaranteed? How to deal with third party components, in particular general-purpose AI models, that are difficult to access during an assessment?
- **Risk assessment and safeguarding**
 - *Operationalization perspective:* What methodologies can be employed to effectively characterize and evaluate potential risks and vulnerabilities, considering both the technical aspects and broader implications? How must AI governance frameworks look like to mitigate those risks efficiently?
 - *Testing tools and implementation perspective:* What strategies or methods can developers employ to select suitable testing or risk mitigation measures tailored to the specific characteristics of their AI systems? What are novel techniques, tools or approaches for quality assurance? How can systematic tests be performed and what guarantees can these tests give? In particular, how can diverse test examples be generated, including corner cases and synthetic data, to enhance the robustness and quality of AI products? How can generative AI be used as part of assessment tools e.g., for generating test cases?
- **Conformity with Regulations**
 - *Operationalization perspective:* How can compliance with the AI Act and upcoming regulations be implemented into AI software and AI systems, particularly in specific use cases, and what steps are required for achieving and maintaining compliance? In other words, how does a trustworthy AIOps framework look like?

¹ Bertrand Braunschweig, Stefan Buijsman, Faïcel Chamroukhi, Fredrik Heintz, Foutse Khomh, Juliette Mattioli, Maximilian Poretschkin. *AITA: AI Trustworthiness Assessment*. In *AI and Ethics 4*, pages 1–3. 2024

² Qinghua Lu, Foutse Khomh, Apostol T. Vassilev, Maximilian Poretschkin. 2nd International Workshop on Responsible AI Engineering (RAIE'24). Foreword to RAIE 2024. In *IEEE/ACM International Workshop on Responsible AI Engineering (RAIE)*, pages 7–7. 2024

- **Business models and practical application of AI assessments**
 - *Operationalization perspective:* What are business models based on AI assessments and what are key success factors for them? How must assessment criteria be formulated and which KPIs are suitable to make AI quality and trustworthiness measurable in specific AI systems? How need AI quality seals be designed and how do they influence consumers' decisions?
- **Infrastructure and automation:**
 - *Testing tools and implementation perspective:* What infrastructure and ecosystem setup is necessary for effective AI assessment and certification, including considerations for data and model access, protection of sensitive information, and interoperability of assessment tools? Which approaches are there to automate the assessment (process) as much as possible?

■ Organizers of the Workshop

Organizing Committee

- Rebekka Göрге, Fraunhofer IAIS, Germany
- Elena Haedecke, Fraunhofer IAIS, University of Bonn, Germany
- Fabian Malms, Fraunhofer IAIS, Germany
- Maximilian Poretschkin, Fraunhofer IAIS, University of Bonn, Germany
- Anna Schmitz, Fraunhofer IAIS, Germany

Program Committee

- Bertrand Braunschweig, Confiance.ai, France
- Lucie Flek, University of Bonn, Lamarr Institute for AI and ML, Germany
- Antoine Gautier, QuantPi, Germany
- Marc Hauer, TÜV AI.Lab, Germany
- Manoj Kahdan, RWTH Aachen, Germany
- Foutse Khomh, Polytechnique Montreal, Canada
- Julia Krämer, Erasmus School of Law in Rotterdam, Netherlands
- Qinghua Lu, CSIRO, Australia
- Jakob Rehof, TU Dortmund, Lamarr Institute for AI and ML, Germany
- Franziska Weindauer, TÜV AI.Lab, Germany
- Stefan Wrobel, University of Bonn, Fraunhofer IAIS, Germany
- Jan Zawadzki, Certif.AI, Germany

Additional Reviewers

- Sujan Gannamaneni, Fraunhofer IAIS, Germany
- Anna Hake, QuantPi, Germany
- Yue Liu, CSIRO, Australia
- Max Losch, QuantPi, Germany
- Michael Mock, Fraunhofer IAIS, Germany
- Mahesh Chandra Mukkamala, QuantPi, Germany
- Maximilian Pintz, Fraunhofer IAIS, University of Bonn, Germany
- Boming Xia, CSIRO, Australia



On Assessing ML Model Robustness: A Methodological Framework

Afef Awadid¹ ✉

IRT SystemX, Palaiseau, France

Boris Robert ✉

IRT Saint Exupéry, Toulouse, France

Abstract

Due to their uncertainty and vulnerability to adversarial attacks, machine learning (ML) models can lead to severe consequences, including the loss of human life, when embedded in safety-critical systems such as autonomous vehicles. Therefore, it is crucial to assess the empirical robustness of such models before integrating them into these systems. ML model robustness refers to the ability of an ML model to be insensitive to input perturbations and maintain its performance. Against this background, the Confiance.ai research program proposes a methodological framework for assessing the empirical robustness of ML models. The framework encompasses methodological processes (guidelines) captured in Capella models, along with a set of supporting tools. This paper aims to provide an overview of this framework and its application in an industrial setting.

2012 ACM Subject Classification Software and its engineering → Software verification and validation

Keywords and phrases ML model robustness, assessment, framework, methodological processes, tools

Digital Object Identifier 10.4230/OASICS.SAIA.2024.1

Category Academic Track

Funding This work has been supported by the French government under the “France 2030” program.

Acknowledgements We would like to extend special thanks to the robustness team of the confluence.ai research program, composed of AI engineers, for producing the deliverables that served as the foundation for this paper.

1 Introduction

Central to Machine Learning (ML) is a “data-driven AI technology that automatically discovers patterns and relationships from large volumes of data using algorithmic models” [27]. In this context, ML enables computers to learn complex statistical patterns from data and make decisions without explicit programming [14]. Moreover, it facilitates tackling “tasks that are too difficult to solve with traditional programming paradigms” [17]. ML has also proven effective in analyzing customer demand, allowing for accurate anticipation and planning of future needs [16]. It is therefore unsurprising that ML has significantly transformed various industries [11]. However, the inherent risks and uncertainties associated with ML technology pose significant challenges, especially when implementing it in safety-critical systems such as autonomous vehicles.

Indeed, the successful deployment of ML models can be constrained by biases present in the training data. If the data used for training is not representative of the actual scenarios that the system will encounter in the real world, the ML model may make inaccurate and erroneous decisions. Such inaccuracies can lead to catastrophic outcomes, such as vehicle

¹ corresponding author



© Afef Awadid and Boris Robert;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 1; pp. 1:1–1:10

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

accidents, posing a significant threat to human life [4]. Furthermore, the complexity of ML-based safety-critical systems makes them susceptible to various adversarial attacks throughout the ML pipeline. This vulnerability stems from the unpredictable and dynamic environments in which these systems operate, where even minor changes in input data can result in serious consequences [24].

To address this challenge, the [Confiance.ai research program](#) has placed particular emphasis on the robustness of ML models by developing a framework to support their assessment. This framework includes methodological guidelines/processes captured in Capella models, as well as a set of tools. It is part of a comprehensive methodology designed to guide the development of trustworthy ML-based systems.

This paper aims to provide an overview of the proposed framework. It, thus seeks to address the following research question: How can we support the assessment of ML models' robustness to ensure their successful deployment in ML-based safety-critical systems?

The rest of the paper is structured as follows. Section 2 introduces the theoretical background of this work. The methodological framework for assessing ML model robustness is presented in Section 3. Section 4 concludes the paper with suggestions for future work.

2 Theoretical Background

2.1 Context and Motivation

Integrating ML techniques into safety-critical systems can promote autonomy and support decision-making processes. However, this increased autonomy may also introduce unpredictability and uncertainty, which could challenge the system's reliability. Accordingly, to ensure the trustworthiness of ML-based safety-critical systems, it is essential to assess the robustness of ML models before deploying them operationally.

Trustworthiness refers to the probability that a system has certain established properties such as robustness and reliability, with higher trustworthiness indicating a greater likelihood that the system will exhibit those properties [5].

Against this backdrop, the Confiance.ai research program has proposed an end-to-end method aimed at guiding the development of ML-based safety-critical systems, with the aim of bolstering the French industry's confidence in these technologies. This method encompasses methodological processes and associated tools, covering the entire lifecycle of ML-based systems, from design to operation. It addresses critical aspects of ML model trustworthiness, such as robustness and explainability. This paper specifically focuses on the part of the end-to-end method dedicated to the evaluation of ML model robustness, highlighting the methodological framework developed for this purpose.

2.2 Related Work

Robustness refers to the ability of an AI system to maintain its level of performance under any circumstances (e.g. external interference or harsh environmental conditions) [2]. It has emerged as a key quality characteristic of trustworthy AI. In this context, [3] emphasizes the importance of robustness as a sub-characteristic of reliability in the quality of AI systems. One reason for this emphasis is that robustness allows AI systems to maintain normal operation and avoid unreasonable safety risks throughout their lifecycle, even in the case of misuse and other unfavorable conditions [1].

Given this, it is not surprising that the robustness assessment of ML models has been attracting significant attention, leading to several approaches in this regard. These approaches can be broadly categorized into two main types: 1) formal verification-based approaches and

2) statistical verification-based approaches. In the first category, robustness verification is formulated as a satisfiability problem, where the goal is to identify the smallest set of inputs that satisfy the robustness condition and result in a bound for local robustness (i.e., the model's ability to maintain its output within specific regions in the input space). Examples of these approaches include stability region verification [9, 7], sensitivity analysis [20], and safety region verification [12, 21, 26].

The second category of approaches relies on statistical verification to assess local robustness. These methods quantify the local robustness of ML models by evaluating the probability of inputs violating or satisfying the robustness condition within the verification region. Such approaches are generally implemented based on input domain sampling, as seen in [8, 13], or within a formal framework (e.g., [23, 25, 6]).

In view of this, existing approaches for ML model robustness assessment tend to address this issue from a purely technical point of view. Consequently, the proposed solutions are tailored to individuals with high technical skills, specifically ML algorithm engineers. Implementing these solutions requires substantial knowledge in ML algorithm engineering and related fields. In contrast, within the Confiance.ai research program, we provide both methodological support (engineering processes/guidelines) and technical support (tools for those guidelines) to assess ML model robustness. Our resulting methodological framework is part of an end-to-end method that covers the entire ML systems engineering lifecycle and aims to guide the development of trustworthy ML systems. Therefore, it is designed to be accessible to various engineering specialists, such as system engineers, ML algorithm engineers, and data engineers.

2.3 An Overview of the End-to-End Methodology

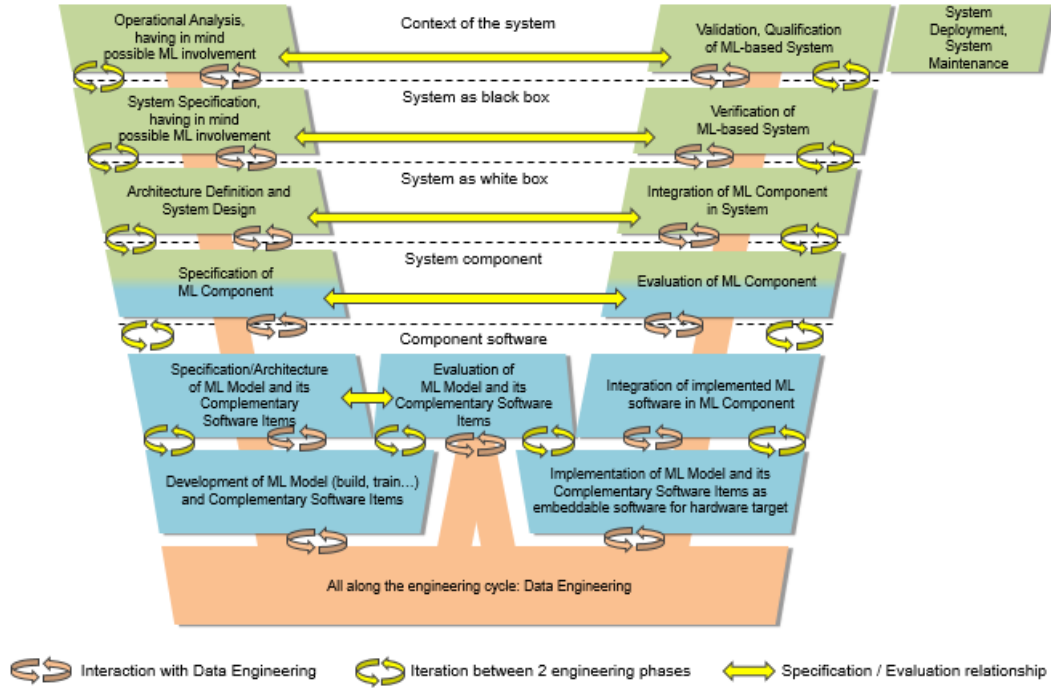
Within the Confiance.ai research program, an end-to-end methodology has been developed by a large and diverse group of experts including industry actors to provide a set of methodological processes/guidelines aimed at assisting in the development of reliable and trustworthy ML-based safety-critical systems. These processes seek to cover the entire lifecycle of ML-based systems (see Figure 1).

Figure 1 outlines the main engineering steps of our methodology, which are defined in line with the ISO/IEC 5338:2023 standard. The latter describes the life cycle of AI systems based on machine learning and heuristic systems. It illustrates the lifecycle of systems engineering that integrates Machine Learning (ML) into the traditional V cycle of system development, creating a customized W cycle for software engineering processes. This W cycle highlights the critical step of evaluating the ML model for reliability at the algorithm level before its implementation at the software level. In this paper, we focus on the engineering activity “Evaluation of ML Model and its Complementary Software Items” in Figure 1.

3 A Methodological Framework for Assessing ML Model Robustness

3.1 Engineering Processes

To evaluate the robustness of ML models, the Confiance.ai research program proposes two engineering/methodological processes. These processes have been captured in Capella models and were developed collaboratively by systems engineers, ML algorithm engineers, and data engineers from various academic and industrial partners of the Confiance.ai research program. The Capella tool was chosen due to its widespread use and familiarity among our multidisciplinary team.



■ **Figure 1** Lifecycle of ML-based systems.

The proposed engineering processes consider two strategies for assessing ML model robustness: 1) robustness testing by sampling and perturbation (also known as empirical robustness) and 2) robustness through formal evaluation (also known as formal robustness). The former involves constructing test datasets with perturbations to evaluate the correctness of a model's output in response to these perturbed inputs. The latter aims to determine if a model is robust within a specific perturbation range and has made significant progress through formal and statistical verification techniques [22].

Both strategies can be combined to assess the robustness of an ML model. However, it is important to note that the formal robustness strategy may not be feasible for certain types of models due to their complexity or lack of formal specifications. In this sense, the adoption of formal verification to evaluate the robustness of ML models depends on certain constraints such as the acceptability of formal proofs, the compatibility of verification tools with the ML model algorithm, and the dimension of the data space. For this reason, the scope of this paper will focus on the robustness testing by sampling and perturbation strategy.

The engineering/methodological process capturing the empirical robustness strategy (i.e., robustness by sampling and perturbation) is presented in Figure 2. A prerequisite of this strategy is that the ML model robustness requirement is expressed as a maximum tolerable deviation of the ML model's behavior in response to a certain intensity of perturbation in the input data.

As shown in Figure 1, the process of evaluating ML model robustness is conducted by the ML algorithm engineer through sampling and perturbation techniques. The engineering activity, titled "Evaluate the robustness of the trained ML model using sampling and perturbation tests," is a sub-activity of the broader task "Evaluate the trained ML model (and its complementary software elements, if necessary)." This indicates that robustness evaluation via sampling and perturbation represents one of several strategies for assessing the robustness of a trained ML model.

Process for [Evaluation of ML Model robustness with sampling and perturbation test]

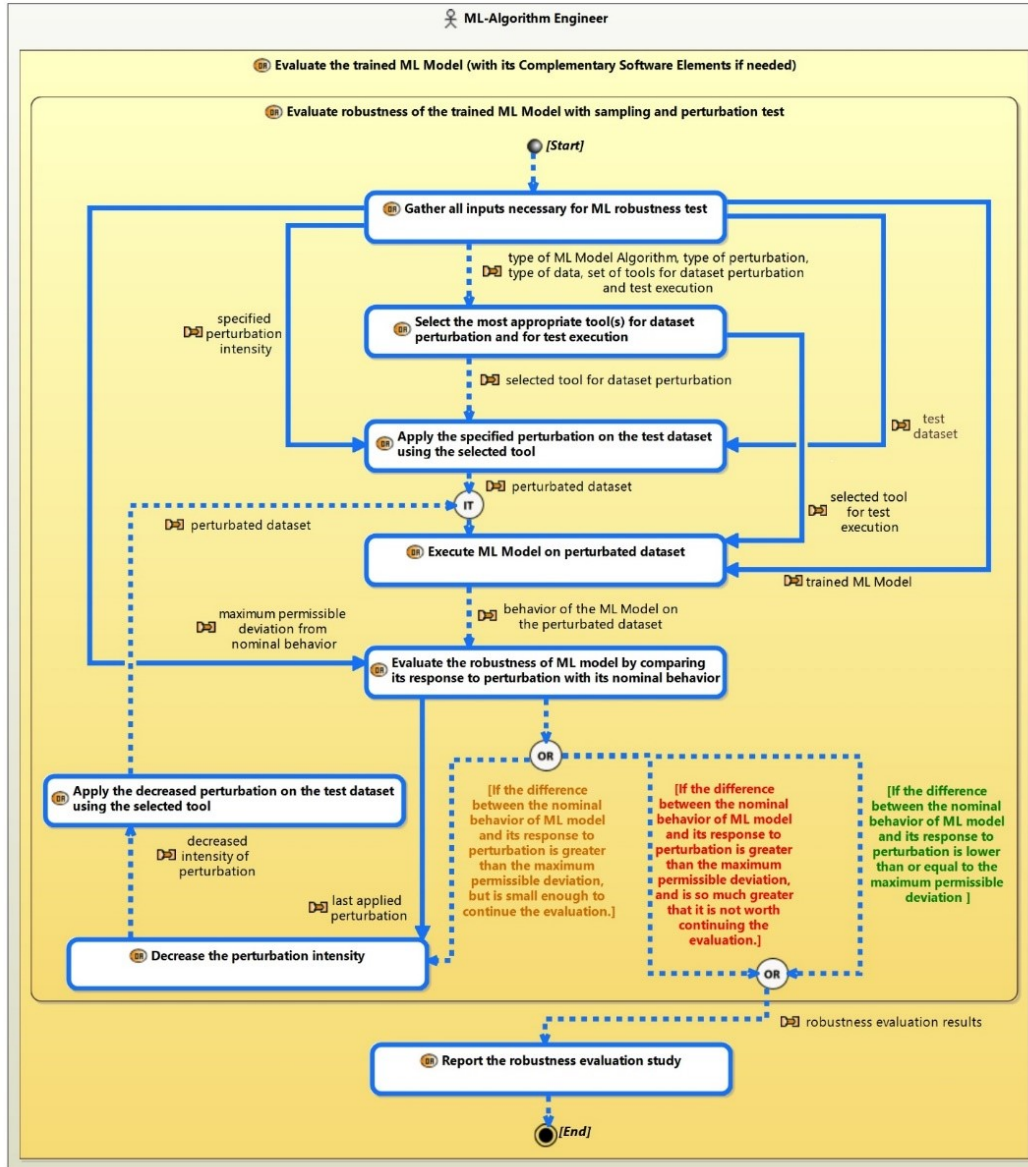


Figure 2 Process for evaluating ML model robustness by sampling and perturbation.

Referring to Figure 2, the process of evaluating ML model robustness according to the empirical robustness strategy includes the following steps:

- Gather all necessary inputs for the ML model robustness test: This initial activity involves collecting all essential information required for testing the robustness of the ML model such as the problem type (e.g. image classification), the data type (e.g. image), the type of perturbation (e.g. variations in image luminosity).
- The second step involves selecting the most appropriate tool for dataset perturbation and test execution from the Confiance.ai set of tools based on the relevant inputs.
- Using the selected tool, the specified perturbation is applied to the test dataset. The tool then executes the test and measures the ML model's performance against predefined KPIs. The resulting behavior of the ML Model is thus captured.
- The robustness of the ML Model is evaluated by comparing this perturbed response with its nominal behavior on the non-perturbed Test Dataset. This involves looking for deviations from expected behavior within permissible limits.
- If the difference between the perturbed response of the ML Model and its nominal behavior is lower than or equal to the given maximum permissible threshold, the robustness evaluation study should report that the initial perturbation level given as input has permitted to reach the target level of robustness. This means that the ML Model robustness requirement is satisfied.
- If such difference is slightly greater than the given maximum permissible threshold, the intensity of the perturbation on the test dataset can be decreased until the difference between the nominal behavior of the ML Model and its response to disturbance is smaller than the specified tolerance value. An alternative approach is to continue the robustness evaluation of the ML model until the perturbation level approaches zero, irrespective of the threshold set for the impact. This method provides a more comprehensive analysis of the model's robustness range.
- If the difference is significantly greater than the maximum permissible deviation, the evaluation study can be reported, as it is not worthwhile to continue the evaluation.

3.2 Supporting Tools

To facilitate the application of the proposed methodological process for evaluating ML model robustness in an industrial setting, the Confiance.ai research program provides a set of supporting tools and components. These components are designed to assess the robustness of trained models against various perturbations. These perturbations can be specific to images, such as Gaussian blur or geometric transformations, and can also include adversarial attacks.

3.2.1 Component 331: Adversarial Attack Characterization Component

This component aims to evaluate the reproducibility level of a decision model by providing its robustness ratio based on specified interest variables, including the variation of their intensity. It relies on the open-source ART-IBM library [19] and evaluates a neural network model against a set of adversarial attacks, such as Projected Gradient Descent (PGD) [15], DeepFool [18], and NewtonFool [10]. The Adversarial Attack Characterization Component has been successfully applied to various use cases.

3.2.2 Component 332: AI Metamorphosis Observer Component (AIMOS)

The [AIMOS component](#) evaluates metamorphic properties in AI models. This toolkit is designed to be agnostic, enabling comparisons across a wide range of model types and use cases. It allows testing various metamorphic properties and transformations over different

value ranges and models, with some visual displays included. AIMOS features several types of attacks, such as Gaussian noise (electrically-induced noise), Poisson noise (thermally-induced noise), Gaussian blur (camera vibration), vertical and horizontal motion blur (camera vibration), pixel, column, and line loss (camera sensor failure), and defocus blur (camera focus variation).

3.2.3 Component 333: Amplification Method for Robustness Evaluation Component

The [Amplification Method for Robustness Evaluation Component](#) assesses the robustness of models using image and time series data by applying amplification methods with noise functions to the dataset. These noise functions are implemented as Python scripts. For image datasets, the provided noise functions include Gaussian blur, horizontal and vertical motion blur, dead pixels, lines and columns, additive Gaussian noise, and multiplicative Poisson noise.

3.2.4 Component 334: Non-overlapping Corruption Benchmarking Component

The [Non-overlapping Corruption Benchmarking Component](#) functions as a tool for assessing the robustness of neural network models using a benchmark of synthetic corruptions. This tool simulates corruptions similar to natural corruptions and tests their impact on image datasets. It evaluates the accuracy drop caused by these corruptions and measures the impact on the model's performance as the severity of the corruption is modified.

3.2.5 Component 335: Time-series Robustness Characterizer Component

The Time-series Robustness Characterizer Component evaluates the robustness of models on time series data using amplification methods with noise functions. These functions are implemented as Python scripts. Specifically, for the time series use case, a frequency-keyed noise function is provided. Each function includes a sample of noise data and a plot describing the evaluation results.

3.2.6 Component 3141: Chiru

The [Chiru component](#) is a tool developed to assess the performance of AI models against input perturbations such as Gaussian noise and Gaussian blur. It provides a graphical interface to visualize the results. Chiru does not inherently support specific model types (e.g., TensorFlow, PyTorch, ONNX, NNET); it is the user's responsibility to add them.

3.2.7 Component 4192: ML watermarking

The [ML watermarking component](#) enables black-box watermarking of ML models to reinforce ownership protection. The objective is to investigate how ML watermarking can protect the ownership rights of model creators and ensure traceability of ML models. This component is not focused on evaluating traditional ML models but rather on assessing watermarked models. The component includes three main categories of attacks against watermarking. The first is watermark removal, where an attacker attempts to remove the watermark from

the model. The second is the ambiguity attack, where an attacker casts doubt on legitimate ownership by providing counterfeit watermarks. The third is the evasion attack, where an attacker tries to evade watermark verification, thereby disabling model identification.

An overview of the compatibility of the presented tools with respect to data type and model type is given in tabular Figures 3 and 4, respectively. Note that the column headers in Figures 3 and 4 correspond, respectively, to the number of components as discussed in 3.2.

Data type \ Component	331	332	333	334	335	3141	4192
Images	✗	✗	✗	✗	✗	✗	✗
Tabular	✗	✗	✗	✗	✗	✗	✗
Time-Series	✗	✗	✗	✗	✗	✗	✗
NLP	✗	✗	✗	✗	✗	✗	✗

■ **Figure 3** Supported data types for each component.

Model type \ Component	331	332	333	334	335	3141	4192
Tensorflow	✗	✗	-	✗	✗	-	✗
PyTorch	✗	✗	-	✗	✗	-	✗
ONNX	✗	✗	-	✗	✗	-	✗
NNET	✗	✗	-	✗	✗	-	✗

■ **Figure 4** Supported model types for each component.

4 Conclusions and Future Work

It has been widely argued that ML model robustness is pivotal to reliable AI systems. Accordingly, assessing this robustness is crucial for the successful deployment of ML models in such systems. With this in mind, in the context of the [Confiance.ai research program](#), we proposed a methodological framework to address this need. The framework encompasses engineering processes and a set of supporting tools, providing both methodological and technical support for empirical and formal robustness assessments. However, for simplicity and clarity, this paper focuses solely on the empirical robustness assessment framework.

The proposed framework is part of an end-to-end method for guiding the development of trustworthy ML systems. This method is accessible via a web application known as the “[Body of Knowledge](#)”, which is currently in development.

The empirical robustness assessment framework presented in this paper (i.e., the process for evaluating ML model robustness by sampling and perturbation and its supporting tools) has been evaluated on industrial use cases such as welding quality inspection, demand forecasting, and visual industrial control. Nevertheless, due to confidentiality reasons on the industrial use cases, the details of the evaluation results cannot be shared publicly.

Yet, an important note to make is that the application of the proposed methodological processes and their associated tools has been highly valuable, as it enabled us to gather feedback from industrial actors and identify areas for improvement. This feedback has already led to actions aimed at enhancing some of the tested tools, such as the [maturation of AIMOS based on industrial feedback](#).




In line with this, as future directions for this research, we plan to pursue two main objectives. First, we will focus on improving the tools that support the engineering processes, using feedback collected from industrial partners. Second, we aim to integrate the developed robustness assessment framework (engineering processes and associated tools) into the “[Body of Knowledge](#)”, so that all results of our research are consolidated into a single comprehensive reference source.

References

- 1 ISO/IEC 22989:2022. Information technology — artificial intelligence — artificial intelligence concepts and terminology, 2022.
- 2 ISO/IEC TR 24029-1. Artificial intelligence (ai)—assessment of the robustness of neural networks—part 1: Overview, 2021.
- 3 ISO/IEC TR 25059. Iso/iec 25059:2023 – systems and software engineering – systems and software quality requirements and evaluation (square) – quality model for ai-based systems, 2023.
- 4 Abraham Chan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan. The fault in our data stars: studying mitigation techniques against faulty training data in machine learning applications. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 163–171. IEEE, 2022. doi:10.1109/DSN53405.2022.00027.
- 5 Gregory Chance, Dhaminda B Abeywickrama, Beckett LeClair, Owen Kerr, and Kerstin Eder. Assessing trustworthiness of autonomous systems. *arXiv preprint arXiv:2305.03411*, 2023. doi:10.48550/arXiv.2305.03411.
- 6 Mahyar Fazlyab, Manfred Morari, and George J Pappas. Probabilistic verification and reachability analysis of neural networks via semidefinite programming. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2726–2731. IEEE, 2019. doi:10.1109/CDC40024.2019.9029310.
- 7 Y Guo. Globally robust stability analysis for stochastic cohen–grossberg neural networks with impulse control and time-varying delays. *Ukrainian Mathematical Journal*, 69(8):1049–106, 2017.
- 8 Chengqiang Huang, Zheng Hu, Xiaowei Huang, and Ke Pei. Statistical certification of acceptable robustness for neural networks. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30*, pages 79–90. Springer, 2021. doi:10.1007/978-3-030-86362-3_7.
- 9 He Huang, Yuzhong Qu, and Han-Xiong Li. Robust stability analysis of switched hopfield neural networks with time-varying delay under uncertainty. *Physics Letters A*, 345(4-6):345–354, 2005.
- 10 Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 262–277, 2017. doi:10.1145/3134600.3134635.
- 11 Mohd Javaid, Abid Haleem, Ibrahim Haleem Khan, and Rajiv Suman. Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, 2(1):15–30, 2023.
- 12 Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24–28, 2017, Proceedings, Part I 30*, pages 97–117. Springer, 2017. doi:10.1007/978-3-319-63387-9_5.
- 13 Natan Levy and Guy Katz. Roma: A method for neural network robustness measurement and assessment. In *International Conference on Neural Information Processing*, pages 92–105. Springer, 2022. doi:10.1007/978-981-99-1639-9_8.

- 14 Ping Li, Fang Xiong, Xibei Huang, and Xiaojun Wen. Construction and optimization of vending machine decision support system based on improved c4. 5 decision tree. *Heliyon*, 10(3), 2024.
- 15 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. [arXiv:1706.06083](#).
- 16 Mohamed-Ilias Mahraz, Loubna Benabbou, and Abdelaziz Berrado. Machine learning in supply chain management: A systematic literature review. *International Journal of Supply and Operations Management*, 9(4):398–416, 2022.
- 17 Dietmar PF Möller. Machine learning and deep learning. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*, pages 347–384. Springer, 2023.
- 18 Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.
- 19 Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Amrith Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.
- 20 Yixin Nie, Yicheng Wang, and Mohit Bansal. Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6867–6874, 2019. doi:10.1609/AAAI.V33I01.33016867.
- 21 Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. *IJCAI-19*, 2019.
- 22 Jie Wang, Jun Ai, Minyan Lu, Haoran Su, Dan Yu, Yutao Zhang, Junda Zhu, and Jingyu Liu. A survey of neural network robustness assessment in image recognition. *arXiv preprint arXiv:2404.08285*, 2024. doi:10.48550/arXiv.2404.08285.
- 23 Stefan Webb, Tom Rainforth, Yee Whye Teh, and M Pawan Kumar. A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*, 2018. [arXiv:1811.07209](#).
- 24 Maurice Weber. *Probabilistic Robustness Guarantees for Machine Learning Systems*. PhD thesis, ETH Zurich, 2023.
- 25 Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. Proven: Verifying robustness of neural networks with a probabilistic approach. In *International Conference on Machine Learning*, pages 6727–6736. PMLR, 2019. URL: <http://proceedings.mlr.press/v97/weng19a.html>.
- 26 Matthew Wicker, Luca Laurenti, Andrea Patane, and Marta Kwiatkowska. Probabilistic safety for bayesian neural networks. In *Conference on uncertainty in artificial intelligence*, pages 1198–1207. PMLR, 2020. URL: <http://proceedings.mlr.press/v124/wicker20a.html>.
- 27 Cong Xu, Wensheng Chen, Mingkuan Lin, Jianli Lu, Yungshiao Chung, Jiahui Zou, Ciliang Yang, et al. Applications and challenges of hybrid artificial intelligence in chip age testing: a comprehensive review. *Journal of Artificial Intelligence Practice*, 6(3):70–75, 2023.

Trustworthy Generative AI for Financial Services

Marc-André Zöller   

GFT Deutschland GmbH, Eschborn, Germany

Anastasiia Iurshina   

GFT Deutschland GmbH, Stuttgart, Germany

Ines Röder 

GFT Deutschland GmbH, Stuttgart, Germany

Abstract

This work introduces *GFT EnterpriseGPT*, a regulatory-compliant, trustworthy generative AI (GenAI) platform tailored for the financial services sector. We discuss the unique challenges of applying GenAI in highly regulated environments. In the financial sector data privacy, ethical considerations, and regulatory compliance are paramount. Our solution addresses these challenges through multi-level safeguards, including robust guardrails, privacy-preserving techniques, and grounding mechanisms. Robust guardrails prevent unsafe inputs and outputs, and privacy-preserving techniques reduce the need for data transmission to third-party providers. In contrast, grounding mechanisms ensure the accuracy and reliability of artificial intelligence (AI) generated content. By incorporating these measures, we propose a path forward for safely harnessing the transformative potential of GenAI in finance, ensuring reliability, transparency, and adherence to ethical and regulatory standards. We demonstrate the practical application of *GFT EnterpriseGPT* within a large-scale financial institution, where it successfully improves operational efficiency and compliance.

2012 ACM Subject Classification Applied computing → Online banking; Applied computing → Document management and text processing; Human-centered computing → Collaborative and social computing; Computing methodologies → Artificial intelligence

Keywords and phrases Generative AI, GenAI, Trustworthy AI, Finance, Guardrails, Grounding

Digital Object Identifier 10.4230/OASICS.SAIA.2024.2

Category Practitioner Track

1 Introduction

The integration of generative AI (GenAI) in the financial sector holds substantial potential, particularly in applications such as credit assessments [4], personalized financial advice [6], customer support [2], investment research [6], or audit assistance for regulatory compliance [12]. For example, GenAI can be used in the compliance office to optimize processes and to increase efficiency. The regulatory framework mandates stringent adherence to rules and policies to mitigate risks such as a violation of data protection. Yet, the complexity and volume of these internal and legal regulations require high effort from experienced auditors to ensure compliance. As an auditing assistant, GenAI can provide comprehensive overviews of existing rules and policies, highlight changes from previous versions, detect conflicting policies, and ensure alignment with internal guidelines. Furthermore, GenAI can identify gaps in the current regulatory framework, thus supporting more robust compliance strategies.

Despite the promising potential, the deployment of GenAI in finance, characterized by its highly regulated environment and serious risk exposure, faces significant challenges. Ensuring that AI-generated content is accurate, reliable, and free from errors in the form of hallucinations [5] is crucial, as inaccuracies can erode trust among users and stakeholders [14]. Moreover, adhering to a strict ethical framework is essential [10]. GenAI systems must be designed and trained to uphold ethical standards, ensuring that user requests are handled



© Marc-André Zöller, Anastasiia Iurshina, and Ines Röder;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 2; pp. 2:1–2:5

OpenAccess Series in Informatics

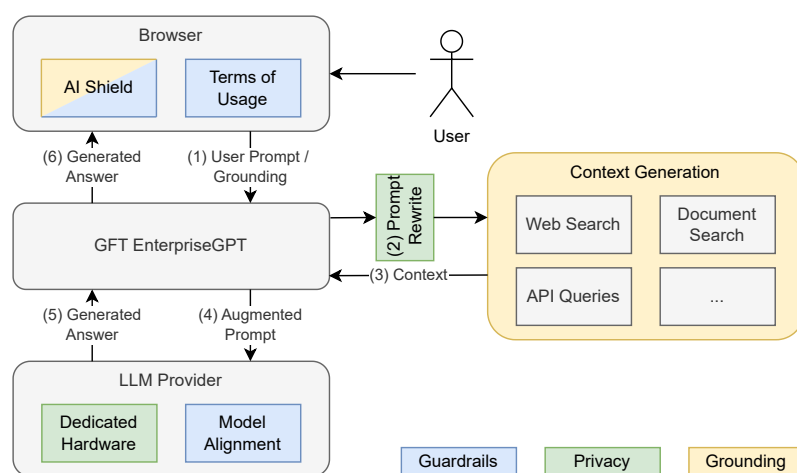


OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

responsibly and that responses are unbiased and transparent. This includes preventing the generation of harmful content or advice that could exploit users. Regulatory compliance is another critical challenge, especially concerning data security. Financial data is highly sensitive, and GenAI systems must comply with stringent regulations like GDPR to maintain trust and avoid legal repercussions, ensuring robust data protection measures are in place [16].

2 Methodology

To address these challenges, our platform, called *GFT EnterpriseGPT*, incorporates several measures to enable trustworthy GenAI. In general, these measures fall into three categories: guardrails, privacy and grounding. Figure 1 provides a high-level overview of *GFT EnterpriseGPT*.



■ **Figure 1** High-level system overview of *GFT EnterpriseGPT*. Displayed in blue, green, and yellow are measures to ensure trustworthy GenAI.

First, the user interacts with their browser to generate a prompt. This prompt is locally analyzed. If the prompt is benign, a request, which is processed by the *GFT EnterpriseGPT* backend, is created. *GFT EnterpriseGPT* can use different tools to augment the user prompt with additional context. Finally, the augmented prompt is forwarded to an large language model (LLM) provider, e.g., a local open-source model or a commercial provider. In the following, we will introduce the separate measures for trustworthy GenAI in more detail.

2.1 Guardrails

Guardrails for foundation models and requests are measures to ensure their safe, ethical, and effective use. They help in mitigating risks, preventing harmful outcomes, and ensuring compliance with legal and ethical standards [15]. In general, a distinction can be made between input and output guardrails. Input guardrails verify that requests that are classified as risky will not enter the LLM model, for example, inquiries on a critical topic such as the construction of weapons. In contrast, output guardrails also check for hallucinations [7]. All foundation model providers implement safety measures [15], like alignment fine-tuning. While this provides a solid basis for ethical GenAI, they are not sufficient in practice as they can not be adapted to specific use cases and can be overcome by malicious actors [3].

Therefore, a client-based prompt analysis, called *GFT AI Shield*, is used to detect, for example, unethical requests or violations of internal policies. This analysis is performed purely inside the client and the prompt under inspection is never forwarded to a commercial GenAI provider to prevent data leakage. Only when no violations, for example, a prompt containing credit card information, were detected, the user prompt is forwarded to an LLM for answer generation. The *GFT AI Shield* uses different methods to detect inadequate content:

- Simple pattern matching is used to detect restricted words
- Named-entity recognition [13] using a local neural network checks if restricted categories are used in the prompt
- Using a plugin interface, customers can provide additional classifiers to detect problematic prompts

If a prompt is flagged by one of the filters, the user is informed why the request is not being processed. All violations against the guardrails are logged in a database. As it is always the case with artificial intelligence (AI) applications, the *GFT AI Shield* can not guarantee the perfect correctness of all predictions. To prevent users from simply switching to public GenAI services, which have a lower security standard, the *GFT AI Shield* is tuned to limit false positive errors.

In general, GenAI applications can either be targeted to an internal or public user base of a company. In the case of an internal application, the employees can receive tailored training for responsible GenAI usage. Furthermore, employers can, to some extent, enforce compliant behavior through internal guidelines and terms of usage agreements. However, just relying on compliant behavior is not sufficient as users may make errors or behave maliciously.

2.2 Privacy

Privacy for GenAI using *GFT EnterpriseGPT* is safeguarded through several key measures. Selecting appropriate licensing is crucial, as it defines the terms under which user-generated prompts, model answers, or datasets for fine-tuning, can be used by the foundation model provider. Even though all model providers guarantee that data is not visible to other users, end-users can not verify this guarantee. Using dedicated hardware, such as PTUs [11] or dedicated servers to host open-source models, helps to prevent unauthorized access and data leaks, providing a controlled environment where sensitive information can be processed without exposure risks. Additionally, the *GFT AI Shield* can also be used for enforcing data privacy. As prompts are directly processed on the user's device rather than on remote servers, the need to transmit personally identifiable information (PII) or confidential information across networks to service providers is eliminated, minimizing the risk of data interception or misuse. Finally, as *GFT EnterpriseGPT* also uses tools accessing the internet, for example using a web search, user prompts are rewritten using GenAI to prevent prompt leakage to another third-party service. Together, these measures create a robust framework for protecting user privacy in the deployment and use of GenAI.

2.3 Grounding

Grounding for GenAI refers to the process of linking their outputs to real-world contexts, facts, or external data sources. This ensures the models' responses are accurate, relevant, and contextually appropriate, enhancing their reliability and applicability in practical scenarios. Retrieval-augmented generation (RAG) [9] is a standard method for grounding [8]. It augments a user prompt with information retrieved from external sources before sending

it to a foundation model. By implementing our own RAG approach instead of using a commercial provider, a fine-tuned solution per use-case can be developed without exposing all data sources to a third party. While this approach can be used to encode domain-specific knowledge into the foundation model, hallucinations can still occur [5]. To ensure the correctness of results, our platform offers the option to search for references for parts of the generated answer. By reusing the retrieval part of retrieval-augmented generation (RAG), we are searching for a context similar to a specific part of the answer selected for grounding. If such a similar context exists, the correctness of the partial answer can be ensured.

3 Use Case

GFT EnterpriseGPT has been successfully implemented at Landesbank Baden-Württemberg (LBBW) [1]. LBBW is a full-service commercial and central bank in Germany. Currently supporting approximately 9,000 employees at LBBW in their daily work, *GFT EnterpriseGPT* has proven its effectiveness in enhancing operational efficiency and compliance in a large-scale financial institution. This success story underscores the practical viability and benefits of trustworthy GenAI in the finance sector following regulatory requirements.

4 Conclusion

In summary, the integration of GenAI in finance offers transformative potential, enabling more efficient and accurate processes in highly regulated environments. By implementing multi-level safeguards addressing key challenges related to trustworthiness, ethical behavior, and regulatory compliance, the full potential of GenAI can be harnessed. The success of our deployment at LBBW illustrates the practical impact and value of combining measures for building trust with GenAI, paving the way for broader adoption of GenAI in finance.

References

- 1 Landesbank Baden-Württemberg. Lbbw startet mit eigener generativer ki-lösung durch, April 2024. [Online; accessed 26-July-2024]. URL: https://www.lbbw.de/artikelseite/pressemitteilung/lbbw-startet-mit-eigener-generativer-ki-loesung-durch_ah8wecz4u8_d.html.
- 2 Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Working Paper 31161, National Bureau of Economic Research, April 2023. doi:10.3386/w31161.
- 3 Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024. doi:10.48550/arXiv.2402.09283.
- 4 J. Galindo and P. Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1):107–143, April 2000. doi:10.1023/A:1008699112516.
- 5 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. doi:10.48550/arXiv.2311.05232.
- 6 Zengyi Huang, Chang Che, Haotian Zheng, and Chen Li. Research on generative artificial intelligence for virtual financial robo-advisor. *Academic Journal of Science and Technology*, 10(1):74–80, March 2024. doi:10.54097/30r2kk80.
- 7 Colin Jarvis. How to implement llm guardrails?, December 2023. [Online; accessed 03-September-2024]. URL: https://cookbook.openai.com/examples/how_to_use_guardrails.

- 8 Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6523–6533. Association for Computing Machinery, 2024. doi:10.1145/3637528.3671467.
- 9 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.
- 10 Bahar Memarian and Tenzin Doleck. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023. doi:10.1016/j.caeai.2023.100152.
- 11 Microsoft. What is provisioned throughput?, May 2024. [Online; accessed 30-July-2024]. URL: <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/provisioned-throughput>.
- 12 Rafet Sifa, Anna Ladi, Maren Pielka, Rajkumar Ramamurthy, Lars Hillebrand, Birgit Kirsch, David Biesner, Robin Stenzel, Thiago Bell, Max Lübbering, Ulrich Nütten, Christian Bauckhage, Ulrich Warning, Benedikt Fürst, Tim Dilmaghani Khameneh, Daniel Thom, Ilgar Huseynov, Roland Kahlert, Jennifer Schlums, Hisham Ismail, Bernd Kliem, and Rüdiger Loitz. Towards automated auditing with machine learning. In *Proceedings of the ACM Symposium on Document Engineering 2019, DocEng '19*, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3342558.3345421.
- 13 Peng Sun, Xuezhen Yang, Xiaobing Zhao, and Zhijuan Wang. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278, 2018. doi:10.1109/IALP.2018.8629225.
- 14 Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, pages 272–283, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3351095.3372834.
- 15 Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness. *arXiv preprint arXiv:2401.00287*, 2023.
- 16 Sreedhar Yalamati. Data privacy, compliance, and security in cloud computing for finance. In *Practical Applications of Data Processing, Algorithms, and Modeling*, pages 127–144. IGI Global, 2024. doi:10.4018/979-8-3693-2909-2.ch010.

EAM Diagrams - A Framework to Systematically Describe AI Systems for Effective AI Risk Assessment

Ronald Schnitzer ✉

Technical University of Munich, Germany
Siemens AG, Munich, Germany

Andreas Hapfelmeier ✉

Siemens AG, Munich, Germany

Sonja Zillner ✉

Technical University of Munich, Germany
Siemens AG, Munich, Germany

Abstract

Artificial Intelligence (AI) is a transformative technology that offers new opportunities across various applications. However, the capabilities of AI systems introduce new risks, which require the adaptation of established risk assessment procedures. A prerequisite for any effective risk assessment is a systematic description of the system under consideration, including its inner workings and application environment. Existing system description methodologies are only partially applicable to complex AI systems, as they either address only parts of the AI system, such as datasets or models, or do not consider AI-specific characteristics at all. In this paper, we present a novel framework called EAM Diagrams for the systematic description of AI systems, gathering all relevant information along the AI life cycle required to support a comprehensive risk assessment. The framework introduces diagrams on three levels, covering the AI system's environment, functional inner workings, and the learning process of integrated Machine Learning (ML) models.

2012 ACM Subject Classification Software and its engineering → System description languages

Keywords and phrases AI system description, AI risk assessment, AI auditability

Digital Object Identifier 10.4230/OASICS.SAIA.2024.3

Category Academic Track

1 Introduction

The integration of Machine Learning (ML) models into complex systems is essential for a wide range of applications, including autonomous vehicles, medical decision support systems, and many more. However, the integration of such ML technologies also introduces new risks which have to be considered, especially in critical applications.

Consequently, novel methods to audit and assess the risks of such systems are crucial. Existing approaches typically analyze ML-based risks by focusing on the models themselves. However, a comprehensive risk assessment should consider the entire system, because although many risks originate from the ML model, their impact and potential mitigation strategies need to be evaluated within the context of the whole system.

To support a meaningful risk assessment, the assessor consequently needs to have an overall understanding of the system under evaluation. For that, a comprehensive description of the entire system, including its application context, functional architecture, and development process, is needed. This AI system description should be easily understandable while providing all relevant information to identify potential risk sources and enable a systematic estimation and evaluation of these risks.



© Ronald Schnitzer, Andreas Hapfelmeier, and Sonja Zillner;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 3; pp. 3:1–3:16

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In this paper, we investigate the requirements for an AI system description to facilitate effective risk assessment. The goal of the paper is to answer the following research question:

- What are the essential components and characteristics of an AI system description that support a comprehensive and effective AI risk assessment?

Based on this analysis, we present the contribution of this paper: A framework that defines the structure and essential elements of a comprehensive description of AI systems.

The paper is structured as follows: In Section 2, we define the scope of the paper, provide theoretical background, and position our paper in the research landscape of related work. In Section 3, we present the results of a requirements analysis for systematically describing AI systems to support effective risk assessment. Subsequently, we introduce a framework to guide practitioners in developing suitable AI system descriptions. In Section 5, we discuss how the in Section 3 identified requirements are met by our framework, including a comparative analysis of related work, and outline the limitations of our approach. Finally, the paper is concluded in Section 6.

2 Theoretical Background and Related Work

This paper studies the requirements for and presents a derived framework to develop an AI system description that supports efficient risk assessment. To enable this investigation, it is crucial to define the following terms: AI system and AI risk assessment. This section provides definitions and important background information on these terms. Furthermore, it highlights significant work related to ours and positions our contribution within the existing research landscape by emphasizing our differentiation from previous work.

2.1 AI Systems

First, a clear definition of the object of consideration is needed. Artificial Intelligence and its subfield, Machine Learning, lack universally accepted definitions in the literature. Recently, the European Union published the EU AI Act [4], setting the rules for developing and operating AI systems within the European Union and potentially having worldwide effects [18]. Due to the significance of the EU AI Act to the AI landscape worldwide, we adhere to the definition provided in the regulation: AI systems are defined as “machine-based systems that are designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infer from the input they receive how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” [4].

Based on this definition, we make two important remarks: First, AI systems are software systems that might be integrated into larger systems potentially consisting of other components, including hardware and sensors. In the following, we refer to anything outside the boundaries of the AI system as its environment. Second, AI systems themselves can be composed of several ML models that are connected in a way to solve a particular objective.

For instance, the perception module of an autonomous vehicle might be an AI system that contains several AI models, such as a computer vision-based pedestrian detector, and a detector to detect danger zones in the environment. Additionally, this AI system might include non-AI-based components, such as the deterministic fusion of several sensors, which relies on conventional software and does not fall under the scope of AI.

The framework for describing the architecture and context of AI systems proposed in Section 4 is designed for AI systems built upon Machine Learning techniques, representing the majority of today’s deployed AI systems. The framework might be generalized to other technologies in the scope of AI, but since this is not the focus of this work we postpone such an analysis to future work.

2.2 Risk Assessment

To understand the term risk assessment, it is essential to consider the broader concept of risk management, independent of the technology. ISO 31000 defines *risk management* as “coordinated activities to direct and control an organization with regard to risk” [7]. Within this framework, risk assessment is a part of risk management that involves “a structured process that identifies how objectives may be affected, and analyses the risk in term of consequences and their probabilities before deciding on whether further treatment is required.” [6].

IEC 31010, specifically, provides a variety of methods for practical risk assessment applications [6]. The choice of an appropriate method depends on several factors, including the type, domain, complexity, and available resources for developing and operating the system under assessment. For instance, *failure mode and effect analysis (FMEA)* is a method that systematically analyzes potential failures of a system and is considered best practice in safety-relevant applications. Notably, FMEA is also part of the end-to-end framework for algorithmic auditing introduced in [14].

Generally, all methods mentioned in [6] rely on some form of system description as input for risk assessment. What these methods have in common is the requirement for a systematic description of the object under assessment.

2.3 Related Work

The significant importance of a structured AI system description is highlighted in the literature on AI assessments and auditing [3, 13, 21]. For example, the AI assessment catalog [13] provides a high-level description of the AI system under assessment, although it primarily focuses on analyzing relevant dimensions of trustworthiness in AI systems rather than on the AI system description itself.

Several established methods systematically describe and document parts of AI systems, such as datasheets for datasets [5] and model cards [11], which document relevant aspects of the data and the model itself. While these works focus on specific elements of AI systems, they do not address the holistic representation of AI systems. However, some works are moving in this direction. For instance, [16] introduce AI Fact Sheets, which address a whole system/service perspective but do not focus on the particular requirements from a risk assessment perspective.

Another important source of related work involves system descriptions not specifically for AI systems but for complex software systems in general. Many of these approaches build upon standardized notations for describing complex software systems, such as UML (Unified Modeling Language) [1] and its evolution, SysML [12]. The integrated system nature is particularly well reflected by the C4 model introduced by Simon Brown [2], which also builds upon the UML notation.

The C4 model introduces diagrams showing different levels of detail, enabling the assessor to understand the broader picture while also allowing detailed inspection where needed. First, the system context diagram depicts the high-level perspective on the system and the environment it interacts with. Then, the container diagram shows the main functional blocks of the system and their interconnections. Third, the component diagram provides abstractions of components within individual containers that should map to real abstractions in the codebase. Finally, the code diagram illustrates relations at the code level, such as relations between implemented methods and functions.

Since the C4 model is designed for software systems in general, it does not account for the specifics introduced by AI systems. The framework we present in this paper is inspired by the structure of the C4 model but is adapted to appropriately reflect AI-specific properties in the context of AI systems.

3 Methodology

By assessing the risks of AI use cases in the form of unstructured interviews with data scientists, we identified five requirements a structured AI system description needs to fulfill to effectively support the assessor in conducting a risk assessment. In the following, we describe these five requirements summarizing our insights from the interviews.

3.1 R1: Flexibility Across Diverse AI Systems

AI systems are utilized in a broad spectrum of applications, each with its unique characteristics and requirements. Therefore, a framework to describe AI systems must be sufficiently flexible to reflect this diversity. A one-size-fits-all template for describing AI systems is impractical, as noted by [16], highlighting the challenge of creating a universally applicable template due to the varying needs of different stakeholders.

However, with focusing on risk assessments and the specific needs of the persons preparing and executing them, it is essential that the AI system description can be tailored to the specific context and requirements of the assessment process. A flexible yet standardized format ensures that the AI system description is applicable to various AI systems, enabling assessors to adapt it to the specific needs of the system being evaluated.

Moreover, this flexibility must not compromise the consistency and comparability of the AI system descriptions. By allowing tailored AI system descriptions within a standardized framework, assessors can achieve a balance between adaptability and the uniformity needed for efficient and effective risk assessment.

3.2 R2: Ensuring Reproducibility

A core feature of an AI system description is reproducibility. An AI system description created by several persons or by the same person several times should be approximately identical to guarantee reproducibility and to enable reproducibility on a risk assessment level. Providing strict and clear guidelines on how to create the specific AI System description is crucial. Reproducibility in AI System descriptions also enhances the efficiency of risk assessors. Variations in AI System descriptions increase the time assessors need to understand the system. Standardized AI System descriptions allow them to quickly grasp the system's details, especially when multiple systems are assessed by the same individual, thereby reducing the time needed for comprehension.

3.3 R3: Integration with Risk Assessment

The goals of a risk assessment process are to identify risk sources, estimate their impact and likelihood, evaluate their significance, and propose mitigation measures. The AI system description must enable the allocation of risk sources to relevant parts of the system.

For instance, 24 AI-specific risk sources that might cause harm have been identified in [17]. Furthermore, it was found that some risk sources manifest at the system level, while others manifest in particular system components such as datasets or the AI models themselves.

Additionally, information about the environment is crucial to estimate and evaluate the severity and probability of certain risks. The AI system description should recognize these factors. For instance, the presence of platform edge doors at railway stations drastically decreases the risk of people being harmed by trains, and are therefore important context information when performing a risk assessment for autonomous trains.

Moreover, the part of a system where the risk source manifests and where a mitigation measure is applied might differ. To enable systematic risk management, the AI system description must cover all these factors comprehensively.

3.4 R4: Grounded in Best Practices

While AI technology introduces many new aspects and risks to consider in risk assessments, the approach to assess these risks should adhere as closely as possible to established procedures. Assessors are already familiar with state-of-the-art methodologies, and unnecessary changes would require additional resources to train new assessors.

Therefore, concepts for describing AI systems should not contradict established procedures. Wherever possible, standardized methods that have proven to be effective, such as using UML or SysML, should be incorporated into the new solution. Additionally, best practices from AI documentation, such as datasheets for datasets [5] or model cards [11], should be recognized.

3.5 R5: Reflecting AI-Specific Characteristics

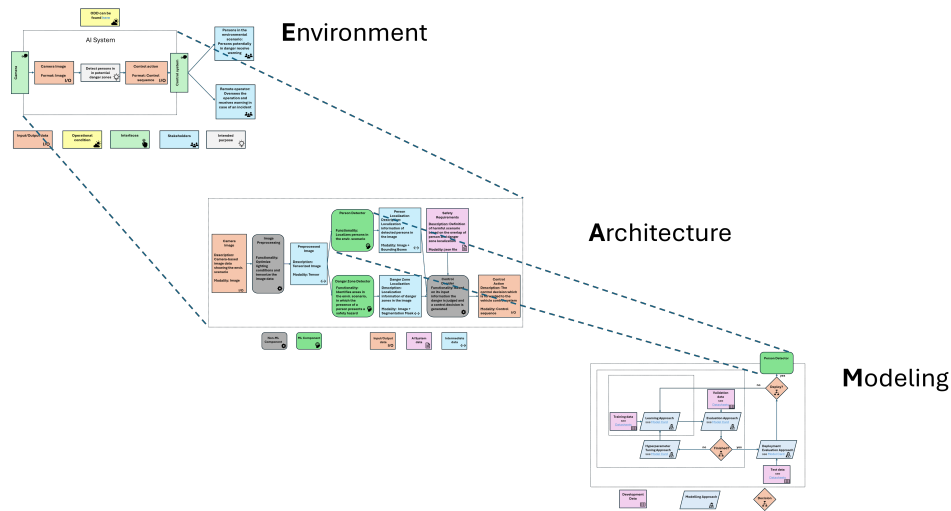
When describing an AI system, it is important to address the aspects ML solutions introduce in comparison to conventional software. It has been intensively discussed in the literature that the use of ML techniques introduces a new set of risks [17, 19, 23]. These ML-specific characteristics include the opaqueness, unpredictability, and complexity of such systems. It is crucial that the framework for describing AI systems reflects AI-specific characteristics to effectively enable AI risk assessment. Another main difference between ML-based solutions and traditional software is how ML models are developed [20]. Especially, the dependence on data sources for the development of ML models is a unique characteristic that must be included in the AI system description. These AI-specific aspects are essential for a comprehensive and accurate risk assessment.

4 EAM Diagrams

Based on the conducted interviews and the requirement analysis, we developed EAM Diagrams, our solution to create an AI system description that is flexible yet comprehensive enough to support efficient risk assessment.

Our concept follows a holistic approach, capturing the interplay between the ML model(s), the AI system and its environment by defining three levels of detail, each represented by its own diagram:

1. The Environment Diagram sets the AI system into context and describes relevant external factors impacting the risk assessment.
2. The Architecture Diagram illustrates the functional inner workings of the AI system, enabling the allocation of different risk sources to the relevant components and representing the operational state of the AI system.
3. The Modeling Diagram relates to the development setup of the ML models integrated into the AI system and covers all aspects of the AI life cycle stages prior to deployment.



■ **Figure 1** Overview of the relation between the different levels of EAM diagram. The three individual diagrams are shown in more detail in Figure 2, Figure 3, and Figure 4.

In the following sections, we describe these three levels in detail as shown in Figure 1, explaining all elements comprising the diagrams. Furthermore, to illustrate the application of EAM Diagrams each level of detail is motivated with a use case as a running example. Figure 2, Figure 3, and Figure 4 show the respective diagrams.

4.1 Running Example

The use case considers a driverless vehicle that operates in an open world and perceives its surroundings via camera input. The person detection system within this vehicle is the AI system under assessment. It aims to detect persons in danger zones, e.g., an area in front of the vehicle and sends respective control sequences to the vehicle's control system to initialize countermeasures in case of danger.

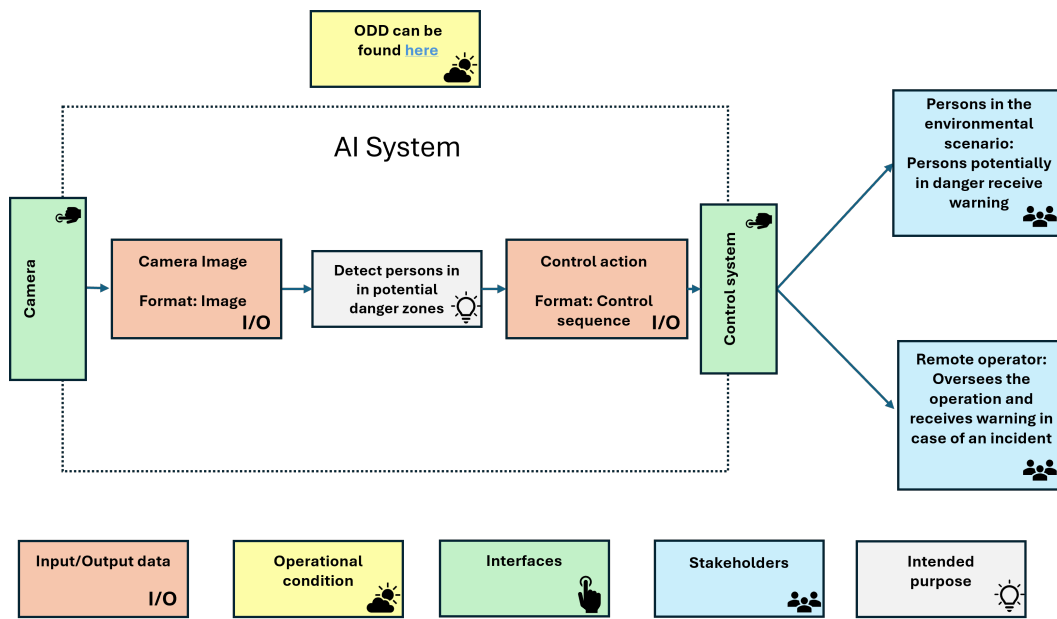
4.2 Environment Diagram

As with any risk assessment process, the first step involves setting the context of the assessment and therefore, the system under consideration [7]. Setting the context for an AI system involves concretely specifying the operating conditions and the system's interfaces to its environment. These aspects form the elements of the Environment Diagram, as depicted in Figure 2 for the running example of a driverless vehicle. In the following, we describe each element, why it is relevant for the risk assessment, and how the right information is included in the diagram.

4.2.1 Intended Purpose

Specifying the intended purpose of an AI system is essential as it not only helps assessors understand the scope of the system but also provides an initial insight into potential risks posed by the system.

The description of the intended purpose should include a textual explanation of the intended use of the AI system by its developer. It is particularly important to describe the intended purpose as concretely as possible, especially if the AI system is based on foundation models, which can be utilized for various downstream tasks. These downstream tasks significantly influence the risks posed by the system.



■ **Figure 2** A schematic Environment Diagram of the driverless vehicle use case.

In case the intended purpose description is too complex for its integration into the diagram, a reference to the document containing the information can be added to ensure accessibility for the assessor.

4.2.2 Operational Conditions

Specifying the operational conditions provides valuable information about the limitations of the system. The level of detail required in documenting these conditions varies depending on the application context. For instance, in a Q&A system, the operational conditions might only state that the system receives user questions of any type, potentially limited to a certain number of input characters. On the other hand, in high-stakes applications, where AI systems operate in complex environments, a more detailed description is necessary. For instance, in the automotive and railway sectors, autonomous vehicles function as high-risk applications in complicated environments. In these sectors, the operational domain is modeled using a so-called operational design domain (ODD), which systematically describes the operating conditions of autonomous systems [8, 10]. This comprehensive description of operational environments is also reflected in the ODD being a crucial input for safety argumentation [22].

In case the description of the operational conditions is too complex to be included in the diagram (e.g., ODDs are typically large documents on their own), the diagram should contain references to the right documents.

4.2.3 Interfaces

AI systems are often part of larger systems. For example, the perception module (an AI system) may be embedded in an autonomous train (the whole system). Issues in the interfaces between the system and its environment might present sources of risk. In such cases, it is crucial to understand the interfaces between the AI system and its environment. If, for example, the AI system receives input from sensors, information about the type of sensor can

be valuable for risk assessment. Another example is a chatbot based on generative AI models, where the interface between the AI system and the human user is typically a graphical user interface (GUI). By making appropriate design choices at the GUI level, such as informing the user that they are interacting with an AI system, risks can be mitigated. Each interface in which the AI system receives input from its environment should be mentioned in the diagram.

4.2.4 Input and Output Data

Understanding the input and output data, as well as their modality, is essential for identifying or excluding certain risks. Specific risk sources may only exist for particular modalities.

For instance, the risk of an AI system hallucinating is associated with AI systems that generate content, such as text, audio, image, or video data [15]. However, this risk is not relevant for an AI system performing a binary classification task. Therefore, understanding the input and output data during operation is crucial information for risk assessment.

Consequently, the modality of the data, as well as a short description, should be denoted in the diagram.

4.2.5 Stakeholder groups

Stakeholders refer to all human persons involved in the operation of the AI system who can be the cause or affected by risks. One potential stakeholder group is the user and its profile, since these are important pieces of information for risk assessment. For instance, if the user group is restricted to specifically trained personnel, the potential for misuse (affecting the risk level of the system) is treated differently compared to a situation where the system is accessible to the general public.

From a risk assessment perspective, understanding the user group helps in identifying reasonably foreseeable misuse of the AI system. This is a crucial part of any risk analysis following ISO/IEC Guide 51 [9] and is required for high-risk systems according to the EU AI Act, Article 9 [4].

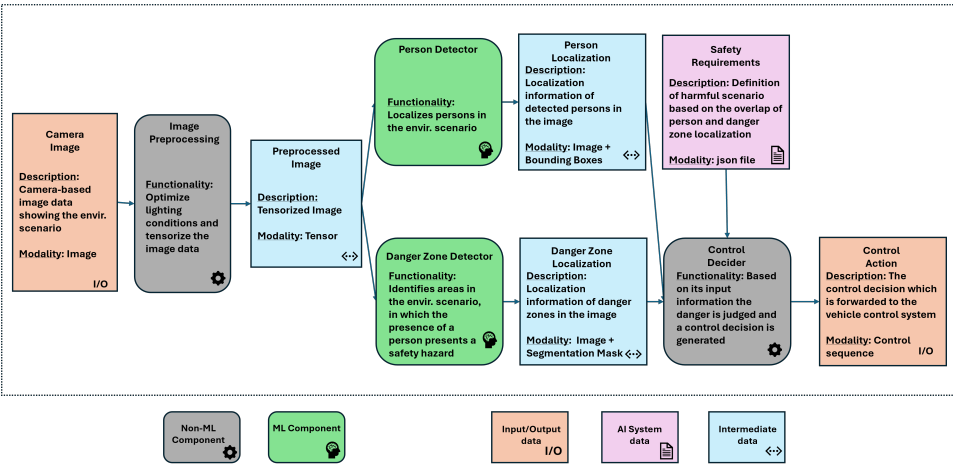
For many systems, the operation is overseen by a human person. For systems to be classified as high-risk by the EU AI Act [4], this is even a mandatory requirement in accordance to Article 14. The presence and level of human oversight provide an indication of the AI system's level of autonomy. Human oversight significantly impacts the risk level and related aspects such as liability and accountability.

Human oversight consists of two key aspects. First, the individuals overseeing the AI system must be aware of the AI system's capabilities and limitations, requiring specific competencies. Second, suitable human-machine interfaces are necessary to enable effective human oversight.

Another important stakeholder group consists of the affected persons. Understanding which groups of persons and how many are affected by the output of the AI system is a crucial factor in evaluating its risk. Analyzing the impact a system failure could have on affected persons is an important part of established risk analysis procedures.

Besides the above-mentioned stakeholder groups, there might be more depending on the specific use case. Of course, if so, they can be added to the diagram as well.

For each stakeholder group, a short description of their role and characteristics relevant to the risk assessment should be provided in the diagram.



■ **Figure 3** A schematic Architecture Diagram of the driverless vehicle use case.

4.3 Architecture Diagram

The purpose of the Architecture Diagram is to describe the inner workings of the AI system. This data flow diagram illustrates the flow of data from input to output, covering all data transformations that occur during the operation of the AI system. It is important to note that the Architecture Diagram does not include the training of the ML models; it only represents the system in its operational state. We introduce two different groups of elements: data elements representing the data itself and component elements representing components that perform data transformations of any type. Together with the data flow (represented as arrows), they form the Architecture Diagram. An application of the Architecture Diagram to the driverless vehicle use case is presented in Figure 3.

4.3.1 Component Elements

Within the component elements of the Architecture Diagram, we distinguish between two types of components. Every component in the data flow diagram should include at least a description of the component’s functionality.

ML Model Components. ML Model Components refer to any ML model created using Machine Learning techniques that, during operation, receives input data to infer output data. This output data is either an output of the AI system or used for further processing within the AI system by other components.

Non-ML Components. In contrast to ML Model Components, non-ML components refer to all components developed without the use of Machine Learning techniques. This includes deterministic pre-processing steps of data, data storage units, and API interfaces. The definition of non-ML components is broad to accommodate a wide range of applications. Necessary information about these components should be detailed in their descriptions.

4.3.2 Data Elements

There are three different types of data elements that are considered. Each data element includes at least a textual description of the underlying data as well as its modalities.

Intermediate Data. Intermediate data refers to any data that has been processed by a component within the AI system's data flow but is not the final output. Intermediate data is the output of one component and the input of another. To avoid inconsistencies, intermediate data is defined as individual data elements.

Input/Output Data. The input and output elements are the same as in the Environment Diagram. However, since they mark the start and end of the operational data flow, they should also be represented in the operational data flow diagram.

AI System Data. AI System data refers to any type of static data that is used by the AI system during inference and is not introduced by the environment. Examples of AI system data include documents that are retrieved by a Retrieval Augmented Generation based Q&A system or (parameter) specifications.

4.3.3 Rules for Constructing the Architecture Diagram

The elements described above form the foundation for the operational data flow diagram. To ensure reproducible results and avoid ambiguity, specific rules must be applied to describe how the different elements are connected. The rules can be summarized as follows:

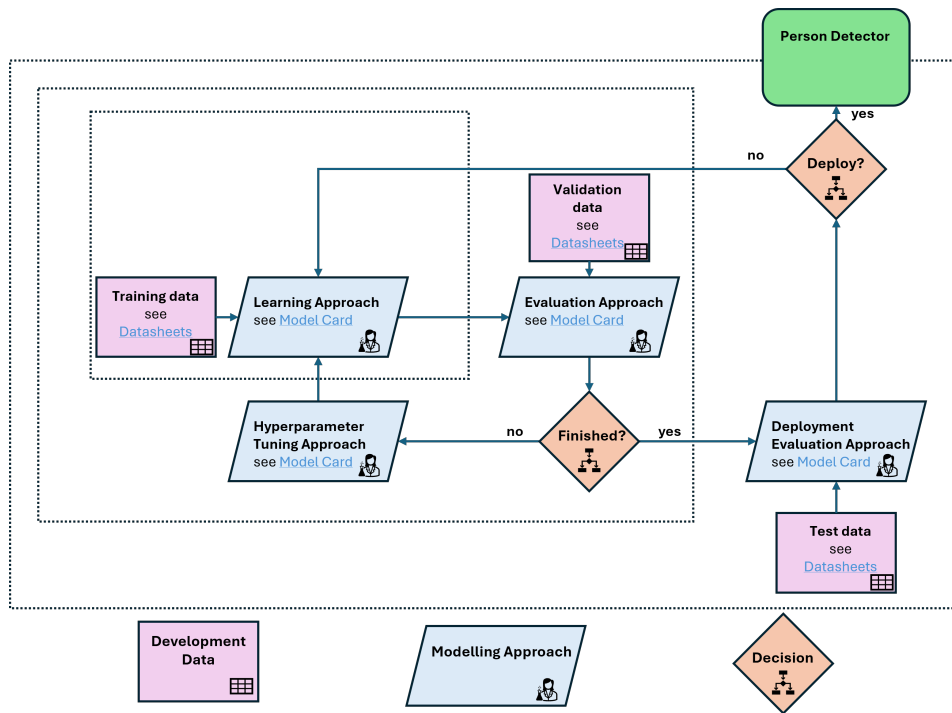
1. Every component element must receive at least one data element as input and produce at least one data element as output.
2. Every intermediate data element must serve as the input to a component element and be the output to another component element.
3. Input data and AI system data elements can not be the output of a component element
4. Output data elements can not be the input to component elements
5. Data elements are only connected with component elements and vice versa.

The Architecture Diagram can also be mathematically represented by a directed graph $G = (V, E)$, where V is a set of nodes and $E \subset V \times V$ is a set of edges, fulfilling the following conditions.

1. $V = V_c \dot{\cup} V_d$, where V_c, V_d represent the set of component and data elements, respectively. $\dot{\cup}$ denotes the disjoint union.
2. $V_d = V_{in} \dot{\cup} V_i \dot{\cup} V_s \dot{\cup} V_{out}$, where $V_{in}, V_i, V_s, V_{out}$ represent input, intermediate, system, and output data, respectively.
3. $e = (v_1, v_2) \in E \Rightarrow e \in (V_d \times V_c) \cup (V_c \times V_d)$
4. $v_c \in V_c \Rightarrow Out(v_c) > 0$, where $Out(v)$ refers to the number of output nodes of the node v .
5. $v_d \in V_{in} \cup V_s \Rightarrow In(v_d) = 0$, where $In(v)$ refers to the number of input nodes to the node v .
6. $v_d \in V_{out} \Rightarrow Out(v_d) = 0$,
7. $v_d \in V_i \Rightarrow In(v_d) > 0 \wedge Out(v_d) > 0$

4.4 Modeling Diagram

One main distinguishing factor between conventional software and ML models is the way they are developed. Conventional software relies on fixed rules specified by the developer for inference, whereas the rules in ML models' are obtained by a learning process using datasets. While ML approaches can solve tasks that conventional software cannot, the mentioned learning process introduces new risks. To address these risks, it is important to understand how the learning process was executed. The ML Modeling Diagram depicts the most important aspects of developing ML models. Figure 4 shows the Modeling Diagram applied to the driverless vehicle use case.



■ **Figure 4** A schematic Modeling Diagram of the driverless vehicle use case showing an exemplary development setup of the Person Detector ML model.

In the following, we detail the relevant elements to create the Modeling Diagram. Note, that arrows in the Modeling Diagram depicts the sequential order in which steps are performed during modeling.

4.4.1 Development Data Elements

There are typically three types of data in the training setup of an ML model: training data, validation data, and test data. Training data is used to adapt the model parameters using the learning algorithm. Validation data is used to evaluate the performance of the trained model, with hyperparameters being tuned based on this evaluation. Once the hyperparameters are fixed, the model undergoes a final evaluation using the test dataset.

Various information about these datasets is relevant for assessing quality and related risk sources, such as quantity and statistical distribution. Use case specific risks may also arise, such as the distribution of ethical characteristics being important in use cases where discrimination is a valid risk. How to comprehensively document datasets in Machine Learning is suggested in datasheets for datasets [5]. If applied, references to the filled datasheets can be added to the ML Modeling Diagram.

4.4.2 Modeling Approach Elements

The Modeling Approach refers to all aspects relating to the actual training process of the ML model. This process includes several design choices by the developer, such as model type, model architecture, hyperparameters, and the learning approach, which may include a loss function and an optimizer.

Another crucial part of any ML model development process is the model evaluation step. Suitable quantification metrics need to be selected to meaningfully describe the relevant performance of the AI system. This also includes specific analyses for certain properties of the AI system, such as accuracy or robustness. Since the choice of quantification metrics and the results can provide valuable insights into how certain risks are controlled, these are relevant information for the risk assessor and are, therefore, reflected in the ML Modeling Diagram.

Note that the aspects of the ML model and ML model evaluation are also addressed by ML model cards [11]. Similar to datasheets, references to these model cards can be included in the ML Modeling Diagram.

4.4.3 Decision Elements

Finally, it is important to highlight that the process of developing an ML model is highly iterative. At various points in the development process, it may be necessary to return to a previous step. This often occurs when evaluation results are not deemed satisfactory. The iterative nature of the development process can be indicated by decisions where a return to a previous step is possible based on specific insights or outcomes.

4.4.4 Rules for Constructing the ML-Modeling Diagram

Since ML developments vary greatly, for instance, based on the type of ML (e.g., supervised, unsupervised, or reinforcement learning) and often involve several iterations, we do not enforce strict rules as we do for the more straightforward Architecture Diagram. However, comprehensive documentation describing the aforementioned aspects should be provided. This documentation is crucial to help assessors identify and evaluate potential risk sources in the ML development setup.

5 Discussion

In this section, we discuss how EAM Diagrams meet the requirements outlined in Section 3 and compare this with related work. Additionally, we address the limitations of our approach.

5.1 Requirement Analysis

In the following we perform a comparative analysis by considering the most related approaches identified in literature. These are the C4 model [2], which introduces also a multi-level diagram concept, and AI Fact Sheets [16]. Although AI Fact Sheets do not introduce a diagram concept, they still introduce a methodology to document various aspects of AI systems, which we consider valid for comparison. An overview of the comparative requirement analysis between the approaches is provided in Table 1.

5.1.1 R1: Flexibility Across Diverse AI Systems

EAM Diagrams are designed with a focus on risk assessments of diverse AI systems. This diversity is considered in two ways: firstly, by defining diagram core elements that are necessary and applicable to various AI systems, and secondly, by ensuring that the properties of these elements are flexible enough to be adjusted to the specific needs of different AI systems while still conveying the necessary information. Consequently, flexibility across diverse AI systems is guaranteed.

The C4 model also provides sufficient flexibility to model any kind of complex software system, including AI systems, using similar arguments. AI Fact Sheets explicitly highlight the diversity of AI systems and their various stakeholders. By providing a general seven-step process for creating AI Fact Sheets for a given use case, they are applicable to a wide range of AI systems.

5.1.2 R2: Ensuring Reproducibility

EAM Diagrams guarantee reproducibility by providing diagram core elements for the AI system description and establishing diagram construction rules where necessary to ensure consistent results.

Similarly, the C4 model introduces specific elements that must be included in the diagrams. AI Fact Sheets also ensure reproducibility by providing explicit guiding questions within each step of the Fact Sheet creation process.

5.1.3 R3: Integration with Risk Assessment

EAM Diagrams are designed to cover the entire AI life cycle: the Modeling Diagram for pre-deployment and the Architecture Diagram for post-deployment. By decomposing complexity into building blocks across different diagrams, while still providing an overview and context of the AI system, assessors can identify and match risk sources, estimate their impact and likelihood, evaluate their significance, and allocate mitigation measures. The visual representation of the AI system provided by EAM Diagrams shows all components and their coordination. This supports a better understanding of the interaction of several risk sources and potential mitigation measures and, therefore, helps in assessing corresponding risks.

The purpose of C4 models is to provide different perspectives on a software system to various stakeholders. However, they do not offer insights into the development process of these systems, which is crucial information for risk assessment due to potential risk sources.

AI Fact Sheets aim to provide holistic documentation of all relevant aspects of an AI system. While they offer general guidelines applicable to any type of stakeholder, they do not focus on specific needs, such as those of risk assessors.

5.1.4 R4: Grounded in Best Practices

EAM Diagrams are grounded in best practices. Their structure is inspired by the C4 model [2] and is compatible with UML and SysML notation, which are established notations in the domain of system architecture descriptions. Furthermore, EAM Diagrams explicitly encourage the integration of best practices in documenting certain aspects of AI systems, such as datasheets for datasets [5] and model cards for model reporting [11].

The C4 model is explicitly compatible with UML. Similarly, AI Fact Sheets build upon state-of-the-art practices in AI documentation and are therefore grounded in best practices.

5.1.5 R5: Reflecting AI-Specific Characteristics

EAM Diagrams are designed to cover the complexity of AI systems and specifically reflect AI-specific characteristics, such as the AI life cycle and the learning process of ML models. To address the learning process, we introduced the Modeling Diagram that explicitly covers these aspects. Additionally, the deployment phase of the AI system is covered by the Environment and Architecture Diagrams. These AI-specific aspects are essential for a comprehensive risk assessment.

■ **Table 1** Overview of the requirement comparison analysis.

Requirement (Section 3)	EAM Diagrams	C4 Model [2]	AI Fact Sheets [16]
R1(Flexibility Across Diverse AI Systems)	✓	✓	✓
R2(Ensuring Reproducibility)	✓	✓	✓
R3(Integration with Risk Assessment)	✓	×	×
R4(Grounded in Best Practices)	✓	✓	✓
R5(Reflecting AI-Specific Characteristics)	✓	×	✓

While the C4 model is well-suited for describing complex software systems, it is not designed to reflect the unique characteristics of AI systems. Consequently, the AI life cycle and the learning process of ML models are not considered in the C4 model. AI Fact Sheets, on the other hand, are specifically designed for documenting relevant aspects of AI systems and therefore reflect AI-specific characteristics.

5.2 Limitations

The EAM Diagrams provide a framework for systematically gathering all relevant information for the risk assessment of AI systems in a diagram format. While the EAM Diagrams claim to be applicable to a wide range of AI systems, they might not be universally applicable to every AI system.

This is due to an implicit assumption about AI systems that was necessary for the structured representation of EAM diagrams. We assumed that the AI life cycle can be strictly divided into development and operation. AI systems that continuously learn after deployment may not be adequately represented by EAM Diagrams. A potential solution would be to integrate a fourth level for such systems depicting the continuous learning plan for such systems. However, since this was not the focus of this research we leave this aspect for future work.

Additionally, the approach assumes that the information required for constructing the AI system diagram is always available. This might not always be the case. When ML models developed or already deployed by third parties are integrated into the AI system, not all necessary information for creating a complete EAM Diagram may be accessible. For instance, if an AI system integrates a deployed version of a foundation model, such as GPT-4 or BART, the integrator might not have access to the information of the training setup or the datasets used in the development process. Notably, this issue is a general problem for all kinds of AI documentation.

6 Conclusion

We presented a framework for systematic creation of an AI system description that collects relevant information to enable efficient risk assessment. If applied, the framework yields three structured diagrams introducing three levels of detail. These represent the AI system's environment, its functional inner workings, and the development process of the integrated Machine Learning model(s). We then demonstrated the effectiveness of the approach by discussing how determined requirements are fulfilled by our approach compared to related work. In future work, we will investigate how EAM Diagrams support compliance with regulations, such as the EU AI Act.

References

- 1 Grady Booch, James Rumbaugh, and Ivar Jacobson. *The unified modeling language user guide*. The Addison-Wesley object technology series. Addison Wesley, 1999.
- 2 Simon Brown. The C4 Model for Software Architecture, June 2018. URL: <https://www.infoq.com/articles/C4-architecture-model/>.
- 3 J Eichler and D Angermeier. Modular Risk Assessment for the Development of Secure Automotive Systems, January 2015.
- 4 European Parliament and Council of the European Union. Artificial intelligence Act, July 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- 5 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. doi:10.48550/arXiv.1803.09010.
- 6 IEC. 31010:2012 - Risk Management - Risk assessment techniques.
- 7 ISO. 31000: Risk Management — Guidelines, 2009.
- 8 ISO. TR 4804:2020 Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation, 2020.
- 9 ISO/IEC. GUIDE 51: Safety aspects: Guidelines for their inclusion in standards, 2014.
- 10 ISO/IEC. 21448:2022 Road vehicles — Safety of the intended functionality, 2022.
- 11 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, January 2019. doi:10.1145/3287560.3287596.
- 12 Object Management Group (OMG). OMG Systems Modeling Language (OMG SysML™), V1.0, 2007.
- 13 Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B Cremers, Dirk Hecker, Sebastian Houben, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss, and Stefan Wrobel. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog, 2023. doi:10.48550/arXiv.2307.03681.
- 14 Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, 2020. arXiv:2001.00973, doi:10.48550/arXiv.2001.00973.
- 15 Vipula Rawte, Amit Sheth, and Amitava Das. A Survey of Hallucination in Large Foundation Models, September 2023. doi:10.48550/arXiv.2309.05922.
- 16 John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. A Methodology for Creating AI FactSheets, June 2020. arXiv:2006.13796, doi:10.48550/arXiv.2006.13796.
- 17 Ronald Schnitzer, Andreas Hapfelmeier, Sven Gaube, and Sonja Zillner. AI hazard management: a framework for the systematic management of root causes for AI risks. In Mina Farmanbar, Maria Tzamtzi, Ajit Kumar Verma, and Antorweep Chakravorty, editors, *Frontiers of artificial intelligence, ethics, and multidisciplinary applications*, pages 359–375, Singapore, 2024. Springer Nature Singapore. doi:10.1007/978-981-99-9836-4_27.
- 18 Charlotte Siegmann and Markus Anderljung. The Brussels Effect and Artificial Intelligence. preprint, Politics and International Relations, October 2022. doi:10.33774/apsa-2022-vx1.
- 19 André Steimers and Moritz Schneider. Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, 19(6):3641, March 2022. doi:10.3390/ijerph19063641.
- 20 Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413, April 2021. doi:10.3390/make3020020.

- 21 Laura Waltersdorfer, Fajar J. Ekaputra, Tomasz Miksa, and Marta Sabou. AuditMAI: Towards An Infrastructure for Continuous AI Auditing, June 2024. doi:10.48550/arXiv.2406.14243.
- 22 Gereon Weiss, Marc Zeller, Hannes Schönhaar, Chrstian Drabek, and Andreas Kreutz. Approach for Argumenting Safety on Basis of an Operational Design Domain. In *2024 IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN)*, 2024. doi:10.1145/3644815.3644944.
- 23 Oliver Willers, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks, January 2020. arXiv:2001.08001, doi:10.48550/arXiv.2001.08001.

Scaling of End-To-End Governance Risk Assessments for AI Systems

Daniel Weimer ✉ 🏠

ceel.ai, Munich, Germany

Andreas Gensch ✉ 🏠

ceel.ai, Munich, Germany

Kilian Koller ✉ 🏠

ceel.ai, Munich, Germany

Abstract

Artificial Intelligence (AI) systems are embedded in a multifaceted environment characterized by intricate technical, legal, and organizational frameworks. To attain a comprehensive understanding of all AI-related risks, it is essential to evaluate both model-specific risks and those associated with the organizational and governance setups. We categorize these as “bottom-up risks” and “top-down risks,” respectively. In this paper, we focus on the expansion and enhancement of a testing and auditing technology stack to identify and manage governance-related risks (“top-down”). These risks emerge from various dimensions, including internal development and decision-making processes, leadership structures, security setups, documentation practices, and more. For auditing governance related risk, we implement a traditional risk management framework and map it to the specifics of AI systems. Our end-to-end (from identification to monitoring) risk management kernel follows these implementation steps:

- Identify
- Collect
- Assess
- Comply
- Monitor

We demonstrate that scaling of such a risk auditing tool requires fundamental aspects. Those aspects include for instance a role-based approach, covering different roles in the development of complex AI systems. Ensuring compliance and secure record-keeping through audit-proof capabilities is also paramount. This ensures that the auditing technology can withstand scrutiny and maintain the integrity of records over time. Another critical aspect is the integrability of the auditing tool within existing risk management and governance infrastructures. This integration is essential to reduce the barriers for companies to comply with current regulatory requirements, such as the EU AI Act [3], and established standards like ISO 42001:2023. Ultimately, we demonstrate that this approach provides a robust technology stack for ensuring that AI systems are developed, utilized and supervised in a manner that is both compliant with regulatory standards and aligned with best practices in risk management and governance.

2012 ACM Subject Classification Computer systems organization

Keywords and phrases AI Governance, Risk Management, AI Assessment

Digital Object Identifier 10.4230/OASICS.SAIA.2024.4

Category Practitioner Track

1 Motivation

The rapid adoption of AI systems across various industries has introduced significant challenges in managing and governing the associated risks. These risks, including algorithmic bias, data privacy concerns, and security vulnerabilities, demand comprehensive risk management frameworks that ensure safety, fairness, and regulatory compliance. However, the increasing



© Daniel Weimer, Andreas Gensch, and Kilian Koller;
licensed under Creative Commons License CC-BY 4.0

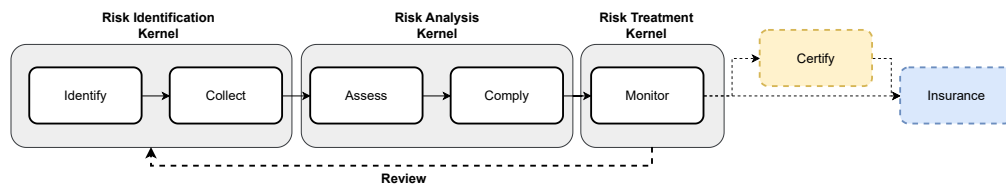
Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 4; pp. 4:1–4:5

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** End-to-end workflow for AI risk management. The circular workflow is based on the general standard for risk management, described in [2].

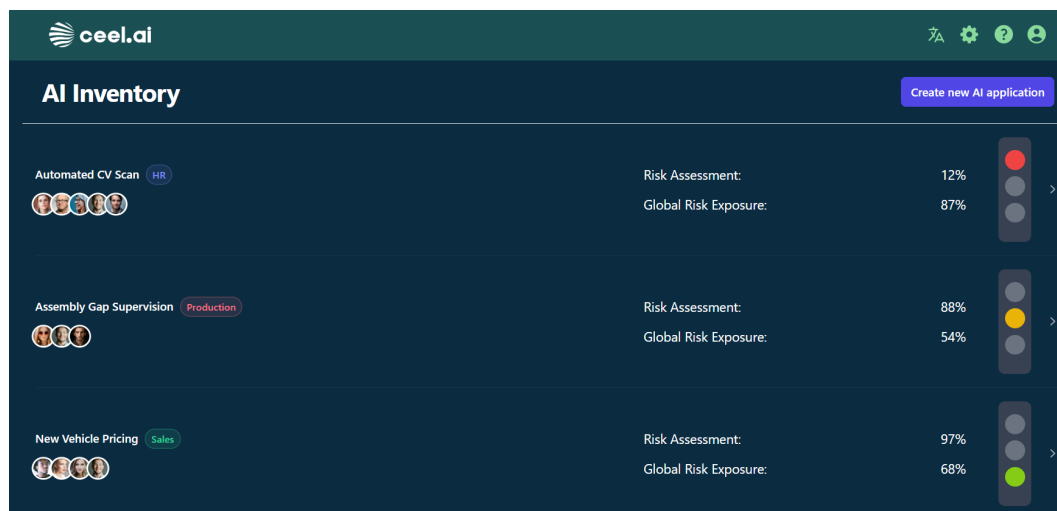
complexity and scale of AI systems make manual risk assessments insufficient to effectively address these issues, necessitating automated solutions for timely and accurate evaluations [4]. Automated AI risk assessments can enhance the consistency, efficiency, and transparency of risk management processes, especially for high-risk applications like healthcare and finance [1]. As AI technologies continue to evolve, the automation of risk management becomes critical to safeguarding ethical and organizational standards and protecting stakeholders from unintended consequences.

To meet those challenges, we have developed an automated risk management technology to enable organizations to fully capture the underlying risk in their AI systems end-to-end. The overall workflow of our technology is visualized in Figure 1 and derived from the general risk management standard described in [2]. The three main pillars of the workflow are risk identification, followed by risk analysis or quantification, while the last step represents measurements and actions towards managing risks in AI systems. The whole workflow is designed as a circular process, involving a constant review. This review is not only limited to changes in potential risks but also to a changing landscape of regulation and standards which might require re-assessments and re-calculation of risks. Section 2 will give more details on the implementation of the described workflow. When implementing the full risk management workflow, certification or insurance of AI systems can be applied in a straight forward way. We represented those two aspects in dotted lines, as only AI systems under a specific risk category might require those aspects.

2 System Design

The design of our automated AI risk assessment tool incorporates a robust audit-proof mechanism, ensuring full traceability and accountability. Central to this system is a write-only architecture that guarantees the immutability of the audit trail, preventing any retroactive alterations or deletions. Every change made within the system is recorded in real-time, capturing essential details such as the nature of the modification, the identity of the user/role responsible, and precise timestamps. This comprehensive audit trail enables a clear reconstruction of the decision-making process and system evolution over time, ensuring transparency and compliance with regulatory standards. By maintaining a secure and tamper-resistant log, the system facilitates complete traceability, allowing auditors to verify compliance and transparency at any given point in time.

Additionally, the system features a comprehensive role model framework, designed to represent various stakeholders within an organization. This ensures that AI risks are assessed from diverse perspectives, including but not limited to technical, legal, and ethical viewpoints, aligning with internal organizational governance and regulatory requirements. The integration of this role-based approach enhances the depth and reliability of the risk assessment, ensuring that decisions are informed by a wide range of expertise and responsibility levels within the organization.

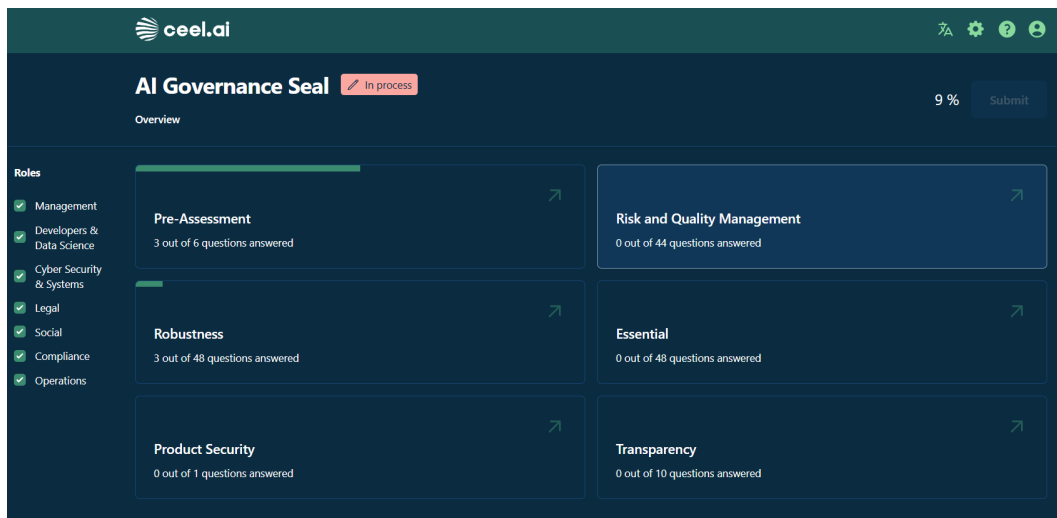


■ **Figure 2** Inventory view in our software suite that collects all AI systems in an organization in one place.

The audit-proof and role based models are key design principles of our automation solution. Following Figure 1, we will describe two aspects of the AI risk management workflow in more detail, namely “Identify” and “Assess”.

- **Identify:** The identification and collection of AI systems within an organization is a fundamental step in the successful assessment and risk quantification process. In our product suite, we provide an AI inventory, as illustrated in Figure 2. This inventory offers a comprehensive, high-level overview of all AI-based systems within the organization, centralized in one accessible location. Implementing an AI inventory enables organizations to track, monitor, and assess their AI assets effectively, which is crucial for managing risks associated with these technologies. Based on our experience working with companies of varying sizes, we have observed that the implementation process tends to be significantly more challenging for large corporations compared to SMEs or startups. This increased complexity arises from the larger number of departments and entities involved in the development of AI systems in large organizations, making coordination and oversight more difficult. As organizations continue to scale, maintaining an accurate and updated AI inventory becomes an essential component of risk governance.
- **Assess:** Each individual AI system collected in the AI inventory must be classified and ranked by risk category to ensure a structured and compliant risk management process. Figure 3 illustrates the assessment kernel implemented within our software solution, where the role-based framework plays a crucial role. Each AI system undergoes a comprehensive risk assessment to classify its risk management category in accordance with the requirements of the EU AI Act, while also providing a deeper understanding of its risk exposure. The software tool is designed to assess AI systems against both public standards, such as ISO and GDPR, as well as custom-defined standards when necessary to meet specific organizational needs. This flexible approach allows organizations to align the risk assessment process with their unique regulatory and operational requirements. As detailed in the system’s overall architecture the assessment is audit-proof, ensuring full transparency and traceability for each AI system under review, thereby facilitating rigorous compliance and accountability across the entire AI lifecycle.

4:4 Scaling of End-To-End Governance Risk Assessments for AI Systems



■ **Figure 3** Assessment view, allowing assessments of AI solutions based on various standards and regulations.

3 Outlook

In this contribution, we have introduced a circular AI risk management workflow and an underlying software solution that automates AI assessments within an end-to-end framework. Central to this framework are traceability and multi-role setups, which, from a system architecture perspective, are essential to meet the requirements of existing and forthcoming regulations and standards. Moreover, we emphasize that risk analysis is not a one-time task but a continuous, circular process requiring the identification of new risks and the ongoing implementation and measurement of regulatory compliance.

Our experience working with organizations of varying sizes reveals significant uncertainty about how to address AI regulation and where to begin. To address this challenge, we recommend a straightforward approach:

- **Identify** all AI systems in your organization (Inventory).
- **Assess** the risk level of these systems in accordance with the EU AI Act.
- **Focus** on high-risk systems and ensure compliance with the relevant regulations.

We strongly believe that our framework, combined with our automation software solution, will simplify the compliance process for organizations striving to meet regulatory requirements. Looking ahead, we anticipate the development of additional standards and technical reports to provide detailed guidance for the successful assessment and certification of AI systems under the AI Act. Future research in AI risk management will prioritize enhancing processes for effective data collection in development and during operations of AI systems, ensuring that collected data are comprehensive, representative, and systematically gathered to support robust risk assessments. Additionally, emphasis will be placed on creating clear compliance frameworks to align with evolving regulations while promoting transparency and accountability. Finally, the establishment of continuous monitoring mechanisms will be crucial for enabling real-time risk detection and adaptive mitigation, ensuring that organizations can respond to the dynamic nature of AI systems and their associated risks.

References

- 1 R. Binns. Fairness in machine learning: Lessons from political philosophy. *CoRR*, abs/1712.03586, 2017. [arXiv:1712.03586](#).
- 2 International Organization for Standardization. Iso 31000:2018 risk management. Technical report, ISO, 2018.
- 3 European Parliament and the Council of the EU. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828, 2024.
- 4 M. U. Scherer. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *European Journal of Risk Regulation*, 29(2):354–400, 2016.

Risk Analysis Technique for the Evaluation of AI Technologies with Respect to Directly and Indirectly Affected Entities

Joachim Iden ✉

TÜV Rheinland Japan Ltd., Osaka, Japan

Felix Zwarg¹ ✉

TÜV Rheinland Industrie Service GmbH, Köln, Germany

Bouthaina Abdou ✉

TÜV Rheinland Industrie Service GmbH, Köln, Germany

Abstract

AI technologies are often described as being transformative to society. In fact, their impact is multifaceted, with both local and global effects which may be of a direct or indirect nature. Effects can stem from both the intended use of the technology and its unintentional side effects. Potentially affected entities include natural or juridical persons, groups of persons, as well as society as a whole, the economy and the natural environment. There are a number of different roles which characterise the relationship with a specific AI technology, including manufacturer, provider, voluntary user, involuntarily affected person, government, regulatory authority, and certification body. For each role, specific properties must be identified and evaluated for relevance, including ethics-related properties like privacy, fairness, human rights and human autonomy as well as engineering-related properties such as performance, reliability, safety and security. As for any other technology, there are identifiable lifecycle phases of the deployment of an AI technology, including specification, design, implementation, operation, maintenance and decommissioning. In this paper we will argue that all of these phases must be considered systematically in order to reveal both direct and indirect costs and effects to allow an objective judgment of a specific AI technology. In the past, costs caused by one party but incurred by another (so-called 'externalities') have often been overlooked or deliberately obscured. Our approach is intended to help remedy this. We therefore discuss possible impact mechanisms represented by keywords such as resources, materials, energy, data, communication, transportation, employment and social interaction in order to identify possible causal paths. For the purpose of the analysis, we distinguish degrees of stakeholder involvement in order to support the identification of those causal paths which are not immediately obvious.

2012 ACM Subject Classification Social and professional topics → Computing / technology policy; Social and professional topics → Computing and business; Software and its engineering → Risk management

Keywords and phrases AI, Risk Analysis, Risk Management, AI assessment

Digital Object Identifier 10.4230/OASICS.SAIA.2024.5

Category Practitioner Track

1 Introduction

AI systems impact humans, societies and the environment through various casual pathways. These effects arise not only through their intended functionalities but also through the limitations of those functionalities (e.g. biases) and the prerequisites for their operation in terms of computing facilities, their construction and material supply needs like electrical

¹ corresponding author



© Joachim Iden, Felix Zwarg, and Bouthaina Abdou;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 5; pp. 5:1–5:6

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

power and water. The term “risk” in the context of this article refers to the possibility of detrimentally affecting entities in various relationships to a planned or deployed AI system. The goal is to present a systematic approach to evaluate possible impacts, which could lead to violations of essential properties for sustainably deploying ethically sound and trustworthy AI systems, without hidden costs or side effects. The Artificial Intelligence Act (AI Act) [2] of the European Union, refers to the following seven principles identified by the AI High Level Expert Group (AI HLEG) as relevant for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal well-being, and accountability. In addition to these principles, the AI Act notably also recognizes energy and environmental sustainability, biodiversity and the rights of the Charter of Fundamental Rights of the European Union as relevant in this context [8, 7].

2 Lifecycle Phases and Corresponding Impacts

Machine learning-based AI systems rely on the collection and processing of large amounts of data, requiring computing equipment often organized in the form of data centers housing great numbers of computing devices with corresponding needs for energy supply and cooling facilities [1, 6]. The lifecycle therefore must take into account the construction of such data centers and the manufacture of their main systems including the computing devices and their supply infrastructure. It is possible to perform an analysis on different levels of granularity. For the current purpose of describing the general approach we distinguish the following phases.

- Data center site construction (physical building, utilities and means of access)
- Data center installation (incl. installation of computing and supply facilities)
- Data center operation (incl. data collection, preparation, model training)
- Data center decommissioning (incl. service discontinuation, building demolition)

Each phase encompasses sub-phases which include dependencies in terms of work and services performed by other businesses, organizations and, ultimately, the human agents involved. Some of this work is directly contracted by the data center operators, while significant parts are hidden in a complex supply and service chain. Indirectly involved work and activities are the manufacture of components for the data center infrastructure and the extraction and processing of resource materials for that purpose. Mining operations for materials like copper, cobalt, lithium and rare earths often have significant environmental impacts, affect human health and agriculture [9]. The demand for specific resource materials may also spawn illegal mining operations where work is performed for minimal income but at high risk to human health and life. Business interests related to mining may further result in direct human rights abuses [4].

Hidden and mostly ignored work involved in data processing includes labelling of data for creating the “ground truth” input data for model training, which are often delivered by an anonymous work force of “volunteer” internet users for minimal pay [3].

Data center operation will provide the basis for companies who deliver derived services utilizing computing facilities and pre-trained models. Such services must be evaluated on their own for their implications. A very successful AI-driven marketing strategy may lead to substantially increased demand for production and shipment of physical items with all pertaining aspects of fair and safe working conditions and environmental impact. Data center decommissioning will involve the aspects of service discontinuation, removal and disposal of infrastructure equipment, and the removal or possible repurposing of the constructed

buildings. Regarding service discontinuation, a relevant question may be whether the service itself is entirely discontinued or whether other facilities are available to provide it instead. Complete service discontinuation needs to be considered with respect to its ripple effects on dependent businesses and general users.

3 Affected Roles / Stakeholders

To address and analyse the risks of AI systems, it is important to understand their impacts related to different domains and costs/benefits for specific stakeholders. In the context of a regulatory technical domain, one can consider the following roles/stakeholders:

- manufacturer
- provider
- user
- workers & employees
- regulatory authority
- certification body

Each of these roles/stakeholders is affected by or is affecting the functionalities of an AI system differently. Therefore, it is important to differentiate the risk analysis based on each role.

4 Degrees of Involvement

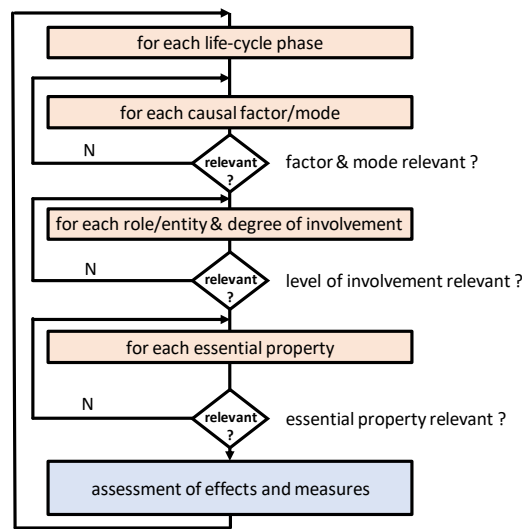
Both the intended functionality and any malfunction of an AI system can exert impact on the various stakeholders. There are different ways to relate to the use of a specific technology. For this reason, and in order to be able to discuss less obvious relationships, we differentiate between several degrees of involvement (refer to Table 1).

The recognition of the 0th degree of involvement allows us to also analyse requirements for sustainability as the environment or society are providing pre-conditions and resources for developing, operating and using any technological system and are in turn affected by such systems in very complex ways.

There is no general or specific relationship between lifecycle phases, stakeholders/roles and degrees of involvement. Instead, the approach is to systematically query a lifecycle process with respect to these aspects and investigate at each phase what entities are instrumental in realizing that specific phase, by what means they contribute to that phase and what are the intended and unintended effects of their contribution.

■ **Table 1** Degrees of Involvement.

Degree	Explanation	Classification
1 st	the main agent or instigator of an activity	intentional and directly affected
2 nd	party intentionally participating in the agent's activity	intentional and in-/directly affected
3 rd	party randomly encountered and not intentionally involved	unintentional and in-/directly affected
n th	party in other location, usually not encountered and whose existence may even be unknown to the agent	unintentional and indirectly affected
0 th	natural environment, society, economy	unintentional and indirectly affected



■ **Figure 1** Risk analysis procedure.

5 Risk Analysis Procedure

The proposed approach to risk analysis has similarities to the concept of causal paths as described in [5] but differs in the detailed application. In causal modeling a causal path is a sequence of events which connects a cause to an effect. This connection can either be direct or involve several intermediate effects. In our approach we aim to construct the cause-effect relations iteratively, for each event repeating the search for modes of causation which will identify further connected events. We consider causal factors or causative media indicated by the following keywords:

- resources,
- materials,
- energy,
- data,
- communication,
- transportation,
- employment,
- social interaction

Performing certain process steps in the overall lifecycle may reveal the need to subcontract services or to procure equipment. These services and the implications regarding the acquired equipment from obtaining the necessary materials, through manufacture to delivery themselves need to be analysed in a similar way, as they contribute to the overall “impact footprint” of the operation. In addition to the causal factors, we also consider the modes of impact exertion, which are expressed through contrastive pairs of terms, for example: consumption - release, gathering - dissemination, demand - supply, improvement - impairment, addition - removal. These causal factors, considered in accordance with their associated modes are evaluated for their impact on the relevant AI properties for affected entities. The examples in Table 2 only show the application of the method in principle. Essential to note is the capacity of the approach to reveal effects, which are often not explicitly stated in the discussion of AI technologies, but which are part of their overall impact and require consideration.

■ **Table 2** Example applications.

causal factor	mode	effect	entity	property	description
physical object (building)	addition	obstruction	environment	environmental integrity	change of microclimate
resource extraction, mining	demand	release of harmful byproducts and substances, dust, smoke, gas	local population	human health	effects on human health due to both direct effects of harmful substances and indirect effects due to diminished possibility for agriculture, forestry, fisheries and tourism
employment	demand	increased employment opportunities	local and global population	fair and safe working conditions	“microwork”, low remuneration

6 Conclusion

We have outlined an approach for systematically investigating the possible impacts of AI technologies at each stage of their deployment lifecycle. These impacts can occur both directly through their intended functionalities and indirectly through the prerequisites for their deployment. Figure 1 outlines the key steps in the analysis. Each traversal of the diagram encounters the step labelled “assessment of effects and measures” and this is where the possible impacts are to be described and documented. In the next steps, effective countermeasures are to be planned to mitigate the detrimental effects which were identified during the analysis.

— **References** —

- 1 Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. *The Datacenter as a Computer: Designing Warehouse-Scale Machines*. Springer International Publishing, 2019. doi:10.1007/978-3-031-01761-2.
- 2 Council of European Union. Council regulation (EU) no 2024/1689, 2024. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2024:1689>.
- 3 P. Hitlin. Research in the crowdsourcing age: A case study, 2016. URL: <http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study>.
- 4 Amnesty International. Democratic republic of the congo: Industrial mining of cobalt and copper for rechargeable batteries is leading to grievous human rights abuses. accessed on 2024-09-16. URL: <https://www.amnesty.org/en/latest/news/2023/09/drc-cobalt-and-copper-mining-for-batteries-leading-to-human-rights-abuses/>.
- 5 Chris Leong, Tim Kelly, and Robert Alexander. Incorporating epistemic uncertainty into the safety assurance of socio-technical systems. In Alex Groce and Stefan Leue, editors, *Proceedings 2nd International Workshop on Causal Reasoning for Embedded and safety-critical Systems Technologies, CREST@ETAPS 2017, Uppsala, Sweden, 29th April 2017*, volume 259 of *EPTCS*, pages 56–71, 2017. doi:10.4204/EPTCS.259.7.

- 6 Peng Li, Jianyi Yang, Mohammad Atiqul Islam, and Shaolei Ren. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models. *ArXiv*, abs/2304.03271, 2023. URL: <https://api.semanticscholar.org/CorpusID:257985349>, doi:10.48550/arXiv.2304.03271.
- 7 Publications Office of the European Union. Charter of fundamental rights of the european union. Technical Report 12012P/TXT, European Union, Brussels, Belgium, October 2012. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>.
- 8 European Commission Publications. Ethics guidelines for trustworthy ai. Technical report, European Commission, Brussels, Belgium, April 2019. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- 9 Nicolás C. Zanetta-Colombo, Tobias Scharnweber, Duncan A. Christie, Carlos A. Manzano, Mario Blersch, Eugenia M. Gayo, Ariel A. Muñoz, Zoë L. Fleming, and Marcus Nüsser. When another one bites the dust: Environmental impact of global copper demand on local communities in the atacama mining hotspot as registered by tree rings. *Science of The Total Environment*, 920:170954, 2024. doi:10.1016/j.scitotenv.2024.170954.

SafeAI-Kit: A Software Toolbox to Evaluate AI Systems with a Focus on Uncertainty Quantification

Dominik Eisl ✉ 🏠

Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Bastian Bernhardt ✉ 🏠

Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Lukas Höhndorf ✉ 🏠

Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Rafal Kulaga ✉ 🏠

Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany

Abstract

In the course of the practitioner track, the IABG toolbox safeAI-kit is presented with a focus on uncertainty quantification in machine learning. The safeAI-kit consists of five sub-modules that provide analyses for performance, robustness, dataset, explainability, and uncertainty. The development of these sub-modules take ongoing standardization activities into account.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases safeAI-kit, Evaluation of AI Systems, Uncertainty Quantification

Digital Object Identifier 10.4230/OASICS.SAIA.2024.6

Category Practitioner Track

1 Introduction

In recent years, Machine Learning (ML) models have become immensely powerful and are used in a wide range of applications and domains. The rise of large language models (LLMs) has led to the popularization of Artificial Intelligence (AI) in society and boosted the already high interest of the industry. Deep learning models are deployed to estimate depth in 2D images [4], to detect abnormalities in medical images [2], or to convert speech to text, often with impressive results. But regardless of the AI task at hand, the number of parameters, or the used architecture, none of the AI models are perfect. Incorrect predictions and mistakes in outputs generated by the AI are inevitable. Given their imperfections, thorough testing and validation of AI models are crucial steps towards deploying them reliably and safely. In the following, we provide insights into the safeAI-kit, which is a toolbox for AI evaluation that offers methods for dataset analysis, performance and robustness evaluation, uncertainty quantification, and explainability examination. In addition, we focus on uncertainty quantification in machine learning.

Over the past years, we have dedicated ourselves to the development of solutions for evaluating and safeguarding AI systems. We combine state-of-the-art AI research, standardization, and regulation with the vast experience of IABG in testing, analyses, and certification processes. By actively participating in AI standardization committees on national, European, and international levels, we support the development of new standards and guidelines to increase the benefits of AI systems.



© Dominik Eisl, Bastian Bernhardt, Lukas Höhndorf, and Rafal Kulaga;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 6; pp. 6:1–6:3

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 IABG's safeAI-kit

To conduct practical AI assessments, we identified the need of a testing tool to support future conformity assessments. The IABG's safeAI-kit is a software toolbox developed to support the evaluation of AI models and datasets and provides a comprehensive analysis with respect to five dimensions. These encompass dataset analysis, performance and robustness evaluation, uncertainty quantification, and explainability examination, all aiming at providing a thorough understanding of an AI model's behavior and capabilities. The resulting evaluation report presents detailed insights into the model's strengths, weaknesses, and overall capabilities, empowering the implementation of safety measures and streamline future audits.

In addition, IABG contributes to the current development of DIN SPEC 92006 "Artificial Intelligence - Requirements for AI testing tools" and aligns the safeAI-kit development with all the concepts and methodology described therein.

3 Uncertainty Quantification in Machine Learning

The real world is complex, chaotic, dynamically changing, and thus difficult to represent in a training set, from which models gain knowledge. Uncertainty is, therefore, inherent in the model's operation. For humans it is very natural to express uncertainty when faced with a new situation or a difficult question. We use phrases like "maybe", "probably" or "I don't know". Analogously, the goal of Uncertainty Quantification (UQ) in ML is to enable the models to signal whether they are confident about the provided output or, on the contrary, that they "don't know" and are in fact guessing.

Hence, uncertainty quantification is an important building block of AI safety which is also vital for various other ML techniques such as active learning. Calibrated uncertainty measures can improve decision making and trustworthiness during operation and therefore offer a step forward beyond offline performance evaluation. Attention and awareness of UQ is growing in scientific and standardization communities as well as industry. As the field of UQ in ML becomes more technically mature and witnesses growing adoption in mission-critical ML tasks, the importance of UQ for safety and certification of AI systems will rise which will help to establish a strong base for the trustworthiness of AI in our society.

Uncertainty Quantification has also already been investigated in standardization. Initiated by IABG and Fraunhofer IAIS and developed by a consortium of experts, DIN SPEC 92005 "Uncertainty quantification in machine learning" [1] is a standardization document which aims to support stakeholders in adopting UQ in ML. The document defines important terms related to uncertainty and provides an overview of UQ applications, approaches, and properties. The core part of [1] is a set of recommendations and requirements for incorporation of UQ in ML. These guidelines aim to help developers navigate through the field of UQ in ML and support them in ensuring that UQ is applied correctly.

4 Conclusion

As the use of AI in safety-critical applications is increasing, independent assessments of such AI systems are essential, particularly considering strict requirements imposed on them by forthcoming legal frameworks, such as the European AI Act [3]. Our goal with the safeAI-kit is to not only contribute to the testing of AI systems but also to support future AI conformity assessment procedures, in alignment with both, legal frameworks and technical standards. The safeAI-kit is continuously evolving to address the challenges arising from limited practical experience in evaluating AI systems for various uses cases and applications. The adoption of

UQ in ML will play a central role for ensuring that ML models are solving complex tasks in a safe and trustworthy manner. Moreover, other dimensions we address with the safeAI-kit – such as performance, robustness, explainability, and dataset analysis – are equally important and must be considered when assessing AI systems. To conclude, we consider AI assessments and audits as a vital step towards enabling robust, reliable, and trustworthy AI systems.

References

- 1 "DIN SPEC 92005:2024-03, Künstliche Intelligenz - Quantifizierung von Unsicherheiten im Maschinellen Lernen; Text Englisch". Technical report, DIN Media GmbH, Berlin, 2024. URL: <https://www.dinmedia.de/en/technical-rule/din-spec-92005/376619718>.
- 2 Minliang He, Xuming Wang, and Yijun Zhao. A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. *Scientific Reports*, 11, 2021. URL: <https://api.semanticscholar.org/CorpusID:233427277>.
- 3 The European Parliament and the Council of the European Union. "REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (Artificial Intelligence Act)", 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- 4 Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024. [arXiv:2401.10891](https://arxiv.org/abs/2401.10891), doi:10.48550/arXiv.2401.10891.

Towards Trusted AI: A Blueprint for Ethics Assessment in Practice

Christoph Tobias Wirth¹   

Smart Data & Knowledge Services, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany

Mihai Maftai 

Ethics Team, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany

Rosa Esther Martín-Peña

Educational Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Berlin, Germany

Iris Merget

Agents and Simulated Reality, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany

Abstract

The development of AI technologies leaves place for unforeseen ethical challenges. Issues such as bias, lack of transparency and data privacy must be addressed during the design, development, and the deployment stages throughout the lifecycle of AI systems to mitigate their impact on users. Consequently, ensuring that such systems are responsibly built has become a priority for researchers and developers from both public and private sector. As a proposed solution, this paper presents a blueprint for AI ethics assessment. The blueprint provides for AI use cases an adaptable approach which is agnostic to ethics guidelines, regulatory environments, business models, and industry sectors. The blueprint offers an outcomes library of key performance indicators (KPIs) which are guided by a mapping of ethics framework measures to processes and phases defined by the blueprint. The main objectives of the blueprint are to provide an operationalizable process for the responsible development of ethical AI systems, and to enhance public trust needed for broad adoption of trusted AI solutions. In an initial pilot the blueprint for AI ethics assessment is applied to a use case of generative AI in education.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Social and professional topics → Codes of ethics; Human-centered computing → Collaborative and social computing; Applied computing → Arts and humanities

Keywords and phrases Trusted AI, Trustworthy AI, AI Ethics Assessment Framework, AI Quality, AI Ethics, AI Ethics Assessment, AI Lifecycle, Responsible AI, Ethics-By-Design, AI Risk Management, Ethics Impact Assessment, AI Ethics KPIs, Human-Centric AI, Applied Ethics

Digital Object Identifier 10.4230/OASICS.SAIA.2024.7

Category Academic Track

Funding The Federal Ministry of Education and Research (BMBF) is funding the project *Artificial Intelligence for Arts Education (AI4ArtsEd)* within the funding measure Cultural Education in Social Transformations as part of the federal research program for Empirical Educational Research.

Acknowledgements We thank Samantha Morgaine Prange and Lisa-Marie Goltz from the DFKI Ethics Team for their valuable contribution throughout the entire paper writing process.

¹ corresponding author



1 Introduction

Artificial intelligence (AI) holds the promise of transforming our world. However, the development of AI technologies leaves also place for unforeseen ethical challenges. Unethical use of AI can lead to various negative outcomes, such as biases and discrimination, privacy and human rights violations, and unintentional harm.

Furthermore, AI practitioners often possess an abstract and somewhat limited understanding of ethical principles and how to translate them into practice effectively. Although their primary motivation is implementing ethical guidelines or principles within practical designs that meet legal requirements, this does not necessarily ensure that AI products are ethically or socially acceptable. Legal compliance alone does not guarantee that AI technologies align with broader societal values or adequately address ethical concerns.

One argument explaining this phenomenon is that new laws often have an extended lead time and cannot keep up with rapidly changing social norms or values. They are not designed to address or adapt to swift shifts in societal expectations. This gap highlights the need for practical ethics to guide practitioners in *operating in the grey areas* [12]. The concept of the grey area refers to ethical dilemmas that emerge when society repeatedly suffers from poor decisions not addressed by existing legislation. These dilemmas often pressure the legal system to adapt and consider new social realities outside existing legal frameworks.

Examples of unethical AI use include Amazon's recruiting algorithm, which displayed a gender bias favoring male applicants over female ones [25]. Another study revealed that AI-based gender classification technology tends to be less accurate for skin types of darker color [5]. Incidents like these can rapidly undermine public trust in AI models' safety, security, reliability, and ethical standards. Without trust, people may fear that AI systems will produce incorrect, inconsistent, or harmful outcomes.

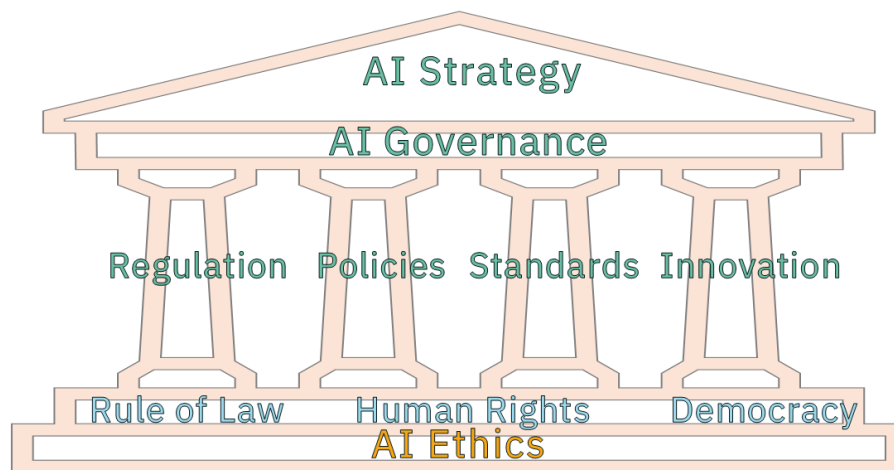
The concept of *Trusted AI* can be explored from multiple distinct perspectives. From the multiplicity of definitions, we understand the term "Trusted AI" as the evaluation of artificial intelligence concerning its reliability and effectiveness in individual applications from the user's perspective, also considering the specific cultural context and values of the community in which the AI system is embedded.

To enhance user trust in AI applications we need to ensure that AI systems are conformant to ethics quality metrics. For this purpose, the German Research Center for Artificial Intelligence (DFKI) Ethics Board has developed a Blueprint for AI Ethics Assessment. In this paper, we present our Ethics-By-Design-based approach aimed at proactively and reactively mitigating the ethical challenges an AI system may encounter during design, development, and deployment.

2 Current global state of AI Ethics implementation

Countries around the world define national AI strategies to leverage the rapid advancement of AI technology. Executing an AI strategy needs governance that includes oversight mechanisms to address risks such as bias, privacy infringement and misuse, but also to build and maintain trust in AI, while at the same time enables AI innovation and research. On international level, the United Nations laid out foundations for the first global architecture for AI governance based on international cooperation [36]. An effective AI governance framework provides a structured approach based on the pillars of regulation, sound AI policy, supporting standards for compliance, and innovation measures. Figure 1 illustrates the building blocks of an AI governance framework. This structure highlights how every element depends on a strong ethical foundation. The AI Strategy represents, in the context of a state, a government's

approach to the development, deployment, and regulation of AI technologies and from a corporate perspective, represents the enterprise AI roadmap. Below, the concept of AI Governance defines the structural support required to operationalize the pillars (regulation, policies, standards, and innovation), aligning them under a unified framework. The four pillars are grounded on a structural basis represented by the foundational aspects of rule of law, human rights, and democracy. At the very bottom, ethics serves as a fundamental grounding, upon which every component and the entire structure as a whole is developed and sustained. This section provides an overview of the current global landscape of AI



■ **Figure 1** Building blocks of a national AI strategy comprise of its governance structure and the functional pillars of regulations, policies, standards and innovation supported by the foundational layer of ethics providing the fundament for rule of law human rights, and democratic values.

ethics, examining how different Digital Empires are responding to the challenges posed by AI. Different regulatory approaches, ethical guidelines, and policy initiatives that have been implemented to ensure that AI technologies are developed and deployed responsibly will be explored. The Digital Empires create a pull effect on other countries in adapting their regulatory approach commonly denoted as Brussels, Beijing, and California effect. The following overview only presents the current point-in-time snapshot of the operationalization potential for AI ethics by selected global digital powers. The choice of geographies is not meant to be biased and presented in alphabetic order.

Africa

The African Union’s (AU) “Continental AI Strategy” prioritizes “economic growth, social progress, and cultural renaissance” [1]. with the help of AI systems. The principles focus on local first and people-centeredness as well as ethics and transparency, inclusion and diversity, human rights and dignity, peace and prosperity, cooperation and integration, and skills development, public awareness and education. This strategy puts forward an Africa-centric and development-oriented and inclusive approach around five focus areas notably: harnessing AI’s benefits, building AI capabilities, minimizing risks, stimulating investment and fostering cooperation. It is part of the AU Agenda 2063 which aims to further peace, prosperity,

7:4 **Towards Trusted AI: A Blueprint for Ethics Assessment in Practice**

self-governance, and international cooperation. The strategy is divided into 5 areas of actions which should be implemented between 2025 and 2030, they are the following: Maximizing AI Benefits, Building Capabilities for AI, Minimizing AI Risks, African Public and Private Sector Investment in AI, and Regional and International Cooperation and Partnerships. Additionally, South Africa has published the “National Artificial Intelligence Policy Framework” [24] and Nigeria its corresponding “National Artificial Intelligence Strategy” [15], both in August 2024.

Canada

In June 2022 the Canadian Government submitted the “Artificial Intelligence and Data Act (AIDA)” [14] under the “Digital Charter Implementation Act” [13], following the “Pan-Canadian AI Strategy” [6] launched in 2017. AIDA adheres to the OECD regulations, the EU AI-Act and the NIST [18] Risk Management Framework reflecting the influence of the Brussels Effect in the Canadian AI strategy, but also the interest in aligning with international standards and ethics requirements to strengthen international/economic relations. AIDA is an addition to existing laws like consumer protection and human rights and will probably come into force in 2025 with administration and enforcement responsibilities lying with the Minister of Innovation, Science, and Industry. In the incipient stages of implementation, the emphasis will be on education, setting up guidelines, and assisting businesses in voluntarily adhering to the new regulation. The government plans to provide sufficient time for the ecosystem to adapt to the new framework before initiating any enforcement action.

China

The National Governance Committee for the New Generation Artificial Intelligence published the “Ethical Norms for the New Generation Artificial Intelligence” [23] in September 2021. The norms for the AI life cycle include fairness, justice, harmony, and security, preventing bias, discrimination, and privacy/information leaks. China has launched the Global AI Governance Initiative (GAIGI) [8] as part of its Belt and Road Initiative, promoting international cooperation in AI governance. Unlike the EU AI Act, China has been regulating specific AI applications individually, such as internet recommendation algorithms, deep synthesis technology, and generative AI. This approach allows China to address specific issues with correspondent rules, building new policy tools and regulatory expertise with each regulation. After the release of ChatGPT the Cyber Space Administration of China (CAC) reacted within 6 months with Draft Measures for Generative AI [37]. China’s AI regulations are designed to be iterative, allowing for quick updates in response to rapid AI developments. The “Interim Administrative Measures for Generative AI Services” [22] exemplify this iterative approach, with the expectation that AI regulation remains highly adaptive.

Europe

In August 2024 the world’s first regulation on AI, the EU AI Act, went into force. This Regulation shall support the EU objective of being a “global leader in the development of secure, trustworthy and ethical AI” [11] and it shall “ensure the protection of ethical principles” [11]. Recognition on the international level of the European legislation reflects the global interest and adaptiveness to the EU regulatory framework, generating the Brussels effect [3]. The AI Act’s binding rules are built on a risk-based approach. However, the implementation of ethics principles for providers and deployers of AI is left on a voluntary basis. The AI Act suggests that for voluntary ethics codes of conduct to be effective, they should be based on clear objectives and key performance indicators to measure the achievement

of those objectives. The AI Act does not explicitly mention that an ethics assessment framework for trustworthy AI must be applied. The AI Act encourages to implement ethics processes in AI system development. In this regard, the EU issued both independently and in collaboration with international bodies multiple ethics principles, guidelines, and assessment frameworks, such as: (i) The High-Level Expert Group on Artificial Intelligence (HLEG) Ethics Guidelines for Trustworthy AI [9], (ii) the Assessment List for Trustworthy Artificial Intelligence (ALTAI) [10], (iii) UNESCO Ethical impact assessment [33].

India

The Indian Government released in 2018 the National Strategy on AI [19]. India focus lies on: healthcare, education, agriculture, smart cities and mobility. Those needs are based on the seven ethics principles: safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency, accountability, and protection and reinforcement of positive human values. These frameworks are not binding, but, for example, the copyright law has been adjusted for AI-generated content. One of the lawsuits against deepfakes was issued after the incident of the Bollywood Actor, Anil Kapoor. His persona had been faked to use for merchandise to earn money. The court agreed with Kapoor since this was a violation of his rights [27]. Furthermore, developments in legislation have been made. The Digital Personal Data Protection Act (DPDPA) was issued in 2023 to ensure the safe usage of personal data to train AI systems [16].

Singapore

Though Singapore does currently not have any binding regulation on AI, the Singaporean government has developed variety of sector-specific and voluntary frameworks to guide the responsible use of AI and to safeguard public interest in AI ethics and governance. In the following two frameworks are introduced, one for financial institutions and the other one for the deployment of generative AI. In 2022 the Monetary Authority of Singapore published assessment methodologies for the fairness, ethics, accountability and transparency (FEAT) principles, to guide the responsible use of AI by financial institutions [17]. The fairness assessment methodology ensures that the AI-assisted decision-making process does not systematically disadvantage individuals or groups of individuals, without appropriate justification. The fairness principle is checked throughout the lifecycle of the AI system's development process based on the key concepts such as selection of personal attributes, types of bias and their mitigation methods, and fairness objectives and their metrics. In 2024 Singapore released the "Model AI Governance Framework for Generative AI" [21] which addresses risks related to Generative AI and provides guidance on practices for safety evaluation of Generative AI models. The framework is based on the core principles of accountability, transparency, fairness, robustness and security and it extends the previous version from 2019 developed for Traditional AI.

U.S.A.

In October 2023 the White House released the Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence. The Biden Administration focuses on eight principles, such as: Safety/Security, Robustness, Reliability, and Repeatability. AI must be standardized and testable before its use to diminish risks. Furthermore, constant monitoring is necessary to ensure ethical development, resilience against misuse, and compliance with Federal laws [31]. The next step is the Blueprint for an AI Bill of Rights, with the

principles: safe and effective systems, algorithmic discrimination protection, data privacy, notice and explanation, and human alternatives, consideration and fallback [28]. Although this is a voluntary framework Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI have offered their commitment [29]. Additionally, 28 healthcare providers and payers have committed to the responsible use of AI in healthcare [30]. The different states can also make their own laws to regulate AI [2]. The Artificial Intelligence Risk Management Framework was published in January 2023 by the National Institute of Standards and Technology (NIST). NIST uses a modified version of the AI lifecycle from the OECD Framework for the Classification of AI systems. After the release of ChatGPT NIST has published the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile in July 2024.

2.1 Implications for AI ethics assessment – The need for a process blueprint

As evidenced by the information presented above, all countries except for the EU AI Act have voluntary regulations or soft-laws when it comes to AI systems. The guidelines often focus on the same principles with security being at the forefront. Privacy and protection are always among the principles, but their understanding differs between countries. As an effect, different court outcomes might appear. In the Indian case mentioned above the court decision was favoring the actor, but in a similar incident in the U.S., when Scarlett Johansson wrote to OpenAI about illegally using her voice, the company stopped the use of her persona, but on a legal level no measures have been taken [26]. This shows that AI governance and ethical frameworks vary across the globe in regard to regional, legal and cultural values, and even more when it comes to strategic interests in shaping digital power.

There are three competing regulatory models, each reflecting a different approach for the digital economy. The United States adopts a market-driven model, focusing on flexible frameworks, China follows a state-driven approach, emphasizing control, security, and social stability in AI development, and the European Union takes a rights-driven stance, prioritizing ethical standards [4]. These three distinct models – market, state, and rights-driven – illustrate that the global landscape of AI ethics is not only a mere reaction of technological advancements but also a manifestation of the underlying political, economic, and cultural dynamics that concretize each region's approach to AI governance.

In summary, a global ethical framework, with the objective to guide the deployment of trusted AI and to promote the responsible use of AI, implies the need of a process blueprint. The blueprint for an AI ethics assessment must fulfill two acceptance criteria. The first criterion refers to its high level of independence, which implies it is agnostic to the underlying regulatory model, to the deployed AI algorithm, to the technology in which the AI model is embedded in, and it is agnostic to the needs of the industry sector or to the business model or scale of business. The second criterion of the blueprint allows for adaptivity to varying comprehension of ethical principles and values. As already been pointed out, the interpretation or choice of ethical principles depends not only on the cultural perspective, but it is also tailored to specific industry needs and it also aims to maximize the space for AI innovation for which most national AI strategies of countries define a leading position. Lastly, the AI ethics process blueprint that fosters a trusted AI ecosystem cannot be static. The blueprint itself requires a review and update process that adapts to advancements in AI.

3 The Blueprint for AI Ethics Assessment in Practice

3.1 Motivation: Blueprint for the entire AI lifecycle

While most AI assessment solutions comprise high-level ethics principles and evaluation tools [7], [20], they miss the practical aspects needed for operationalization in the cycle from idea-to-AIOps deployment. Therefore, our aim is to build a generic AI Ethics Assessment Blueprint for the evaluation of the entire lifecycle of an AI system, from design and development to deployment.

The Blueprint's adaptable framework integrates ethical principles and their associated assessment tools as inputs, leading to a materiality analysis of the AI system. To achieve our goal, we utilized the UNESCO Ethics Principles [32] and the UNESCO Ethical Impact Assessment Tool [33]. We chose the UNESCO ethics framework for two reasons, first, it is congruent with the EU definition of trustworthy AI and, second, it is a global reference standard, adopted by all 193 UNESCO member states in November 2021. An overview of the UNESCO Ethics Principles is provided in appendix A.

The Blueprint for AI Ethics Assessment serves as a facilitator, ensuring that the development process and lifecycle of an AI system are supported rather than constrained. It is designed to enhance and ease the ethical evaluation process, but also to support the ethical and responsible design, development, and deployment of AI systems, providing a structured approach that does not hinder the AI system different lifecycle phases.

3.2 Key Requirements: Successful Implementation of AI Ethics Assessment

In accomplishing operationalization, an AI ethics assessment framework must contain at least the following three components: (i) high-level ethics principles, (ii) an ethics assessment tool corresponding to the ethics principles, and (iii) a set of evaluation measures relating to key performance indicators (KPIs).

Ethics metrics or their defined thresholds provide an important instrument in the decision-making process, for example in selecting mitigation strategies as part of the results of an ethics assessment. Without outcome-driven ethics metrics along the AI lifecycle pathway the operationalization of an ethics assessment framework remains a challenging milestone. To solve the challenge, we propose to develop a phased approach which is described in the next section.

3.3 Structure: The need for a phased approach aligned to the AI lifecycle

The decision to implement a five-phase process in the Blueprint for AI Ethics Assessment is rooted in the need to establish a structured approach to addressing ethical challenges throughout the entire lifecycle of an AI system. This phased approach was chosen to ensure that ethics are not treated as an afterthought or a box-ticking exercise but are an integrated part of AI development and deployment. Because AI technologies present complex and multifaceted ethical dilemmas that require ongoing, context-sensitive assessment, a single-stage process would be insufficient to capture the nuanced and evolving nature of the ethics issues.

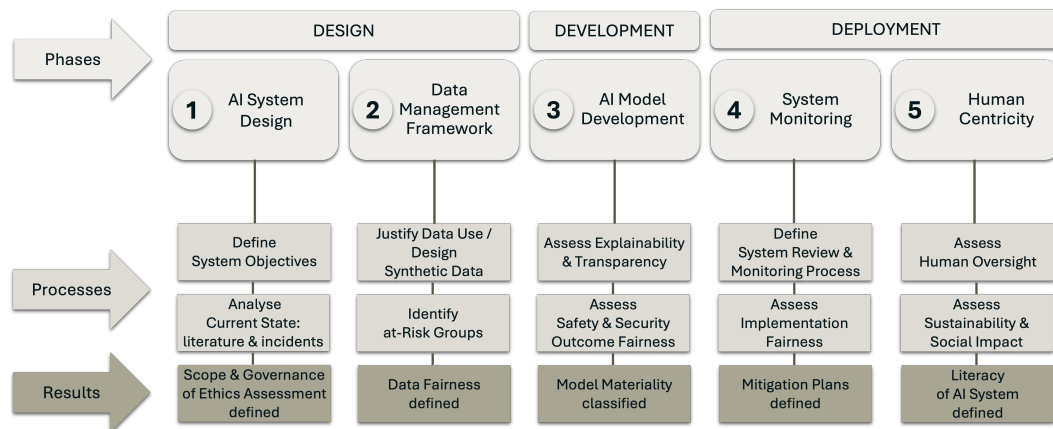
The five phases are based on two motivational drivers: first, to reflect the ethics principles and second, to incorporate the technicalities of the software-engineering needs. Furthermore, focusing on concrete phases such as system design, data management, model development,

system monitoring, and human-centric evaluation, the Blueprint can ensure that ethical issues like fairness, transparency, and accountability are assessed in relation to the actual system development process.

Each of the five phases corresponds to a distinct aspect of the AI system's development, from design to human-centricity evaluation, allowing for a step-by-step integration of ethics in an iterative manner, with each phase building upon the previous one. The rationale behind dividing the process into five phases is to break down the complexities of AI ethics into manageable components, each targeting specific risks and challenges that might emerge at different stages of the AI lifecycle. This phased approach enables continuous feedback loops, ensuring that ethical compliance is not static but evolves alongside the system itself, thus creating a more dynamic and responsive framework. A detailed description of each phase will be presented below.

3.4 Specification: Detailing out the five phases of the Blueprint

Our framework addresses the three main first level stages of an AI system lifecycle - the system design stage, the development stage, and the deployment stage. On the second level, we define five phases where each phase entails two processes on the third level and one result as outcome. Every process includes methodologies and practices aimed at addressing the specific needs and challenges of its corresponding stage within the AI system lifecycle. A schematic of the Blueprint for AI Ethics Assessment in Practice is shown in figure 2.



■ **Figure 2** Schematic of the AI Ethics Assessment Framework for the responsible design and achievement of Trusted AI.

Below, the detailing out the five phases of the Blueprint are presented:

The 1st phase. AI System Design starts with the first process, Define System Objectives, during which the stakeholders define together with the ethics board the objectives and purpose of the AI system. In the second process, Analyse Current State: literature and incidents, an examination of existing literature and relevant incidents is performed to identify potential ethical challenges and best practices to address associated risks. As a result, the scope and governance of the ethics assessment are defined.

The 2nd phase. Data Management Framework focuses in the first process, Justify Data Use / Design Synthetic Data, on the procedural assessment of data use and (when applicable) on the design of synthetic data with the aim to enhance privacy protection and to facilitate

controlled experimentation without compromising sensitive information. The second process, Identify At-Risk Groups, analyzes the prospective data-related ethical issues and it ensures that social justice and equity is promoted. This includes addressing the needs of diverse age groups, cultural and linguistic communities, persons with disabilities, gender diversity, and disadvantaged, marginalized, and vulnerable individuals.

The 3rd phase. AI Model Development starts with the first process, Explainability and Transparency Assessment, which is designed to monitor system outputs and AI-supported decisions to ensure that outputs are explainable, transparent, and aligned with ethics guidelines and stakeholder expectations. This process shall identify checkpoints for feedback collection and continuous monitoring to align the system with human needs and ethical standards. The second process, Assess Safety and Security, selects measures to ensure safety and security of the AI system within a set of defined categories, such as data safety, system robustness, functionality, and detection of potential vulnerabilities of (cyber)security.

The 4th phase. System Monitoring focuses on the first process, Define System Review and Monitoring Process, on the development of a protocol for system evaluation. System monitoring is designed as a continuous process and its goal is to ensure that the AI system operates ethically throughout the whole AI system lifecycle. The second process, Assessment of Implementation Fairness, will evaluate if social justice, fairness and non-discrimination are safeguarded in all AI system structures and layers, for example data intake, algorithmic processing and decision-making.

The 5th phase. Human Centricity starts with the first process, Assessment of Human Oversight, to ensure that the AI system includes different dimensions of oversight which include developer oversight, public oversight, user oversight, and reviewer oversight. During this process, a documented procedure for collecting and analyzing user feedback shall be developed to detect and address ethical challenges in all AI system lifecycle stages. The second process, Assessment of the Sustainability and Social Impact, ensures the continuous assessment of human, social, cultural, and environmental impacts of the AI system. This process shall identify sustainable practices which in turn can address adverse effects on societal and environmental levels.

3.5 Applicability of the AI Ethics Assessment Blueprint

The blueprint is designed to enable the operationalization of AI ethics assessment. The applicability of the blueprint is manifold. It addresses AI systems working alongside human subjects, for example in robotic-assistance or in AI-supported decision-making. It assesses impact along the AI supply chain where downstream AI-driven products or solutions are built around a (generative) AI model offered by an upstream provider. To allow for diverse applications to be assessable, we established a harmonized terminology by mapping assessment questions of a chosen ethics framework to our phases and processes of the AI ethics assessment blueprint. Our approach focused initially on measures of the UNESCO ethical impact assessment with about 160 questions or measures, but it can easily be extended to other frameworks, for example to the European Commission's Assessment List for Trustworthy AI (ALTAI). The mapping of measures is visualized in the dendrogram in figure 3. Answers to the ethics framework assessment questions will then contribute to the outcomes of the blueprint in practice. An effective and timely assessment will require a screening and application or use case specific selection of framework questions. It is not required to answer all questions

to summarize in a meaningful outcomes report. To guide the screening process of questions for relevant outcomes a definition of the outcomes measures of the blueprint is presented in figure 3.

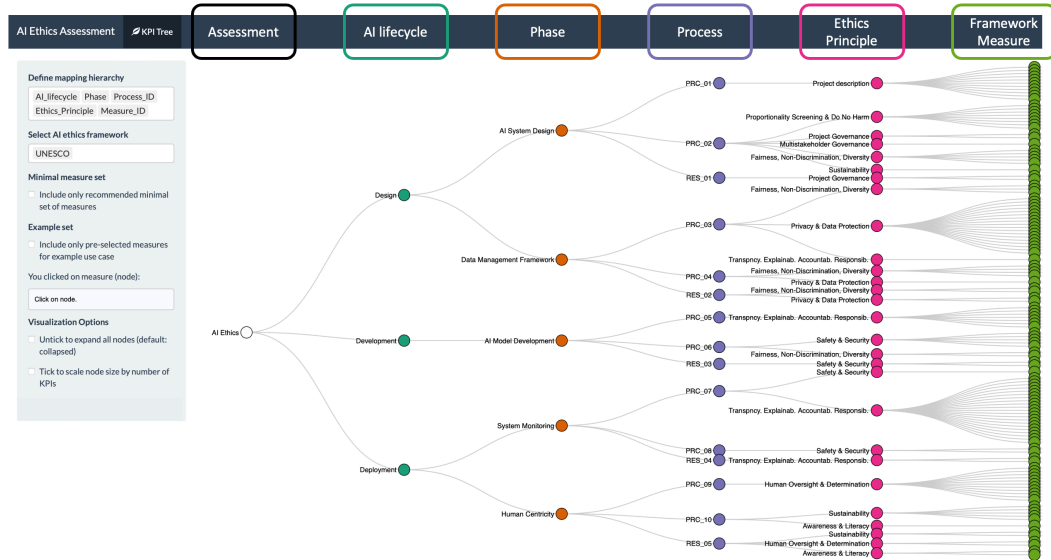


Figure 3 Selection tool for AI ethics assessment blueprint with hierarchical mapping of ethics framework measures to ethics principles, processes, phases and stages of the AI lifecycle.

3.6 Outcomes catalogue for the AI Ethics Assessment Blueprint

Given the diversity of AI use cases the blueprint can address we cannot define application specific output measures and instead we propose an outcomes category for each of the 5 phases. For each outcomes category we provide a set of selectable outcome measures which we denote as AI Ethics key performance indicators (KPIs). The AI Ethics KPIs will then ensure the responsible design, development, and deployment of AI systems. The following outcomes catalogue is exemplary and has no intention of being complete.

3.7 AI Ethics KPIs for outcomes category of phase 1: “Scope and Governance of Ethics Assessment defined”

- **Define governance:** Assemble an Ethics Board, with roles, responsibilities, system objectives, and accountability structures defined within the first few months of project initiation.
- **Identify potential incidents:** Analyze current literature to identify potential incidents, and related proposed mitigation measures for the specific use case.
- **Define review process:** Establish regular review mechanism for ethics clearance by process of multi-stakeholder collaboration including ethics advisors.
- **Define scope:** Select AI system features that must undergo ethical screening by the Ethics Board.

3.8 AI Ethics KPIs for outcomes category of phase 2: “Data Fairness defined”

- **Identify at-risk-groups:** Identify at-risk groups which may be systematically disadvantaged by the AI system.
- **Define fairness:** Select fairness objectives and associated fairness metrics to measure consequences of biased or unfair data on model outputs with respect to harms and benefits which (at-risk) individuals may receive by use of the AI system.
- **Implement bias detection:** Implement bias detection processes throughout the AI lifecycle at defined bias checkpoints based on selected fairness metrics. Definition of processes for mitigation of bias present in data or outputs that impact fairness of the AI system.
- **Implement fairness audit:** Define regular screening and data audits to ensure compliance with fairness data guidelines and ethics principles.

3.9 AI Ethics KPIs for outcomes category of phase 3: “Model Materiality classified”

- **Assess transparency:** Document performance and uncertainty of the AI model with respect to fairness objectives. Justify personal attributes in the data that are used for fairness assessment.
- **Assess explainability:** Conduct explainability assessment of the AI system to ensure that the system’s potential decision-influencing processes are clear and understandable to stakeholders of the system and to users and addresses of the system’s outcome. Explainability assessments are of paramount importance for decision-assist systems. Here the focus is on detecting decision boundaries and deriving concrete recommendations for actions in gray area situations or for high-stakes decisions.
- **Assess safety and security:** Conduct a safety and security evaluation of the AI system to ensure all identified risks to fairness and operational safety are documented and classified by materiality prior to deployment.
- **Identify AI materiality:** Document all risks associated with the AI model’s materiality and categorize all impacts with respect to severity and likelihood. Identify mitigation strategies for each risk.
- **Update materiality classification:** Define process for post-hoc assessment or audit of the AI model’s materiality classification. Flag any newly observed risks and resolve in continuous system improvement initiatives.

3.10 AI Ethics KPIs for outcomes category of phase 4: “Mitigation Plans defined”

- **Define system monitoring:** Define system monitoring and review process to detect abnormal operation of the AI system. Address all identified ethical, operational, and security risks so that potential system impacts are aligned with fairness objectives.
- **Measure incident metrics:** Track incident response metrics so that monitoring enables fast time to detect (TTD) and fast time to resolve (TTR).
- **Define mitigation plan:** Define fallback or mitigation plan in case of trigger events from system monitoring or review.

- **Update mitigation strategies:** Evaluate effectiveness of mitigation plans by measuring incidents after a post-implementation phase. Align updates of mitigation measures with new risks or evolving model behaviors.

3.11 AI Ethics KPIs for outcomes category of phase 5: “Literacy of AI System defined”

- **Guide safe and responsible use:** Develop operationalization guidelines for the AI system. Implement AI literacy and ethics awareness program to ensure that all stakeholders understand how to use and interact with the AI system considering ethics, and system limitations. Ensure that users of a collaborative AI system understand the embodied ethics under normal operation and ethical boundaries. Assess regularly (by user feedback or questionnaires) stakeholders’ ability to exercise human oversight over the AI system. Train developers and system users in recognizing and mitigating ethical risks.
- **Evaluate human centeredness:** Analyze by regular post-deployment reviews that human-AI interaction and human oversight remain effective. These reviews will track how effectively the system supports AI-assisted decision-making.
- **Establish AI training for professional development:** Ensure regular updates of the AI literacy and sustainability training. Adjust training programs based on explaining observed versus expected outcomes, on system improvements, user and stakeholder feedback, and on advancement in state-of-art and energy-efficient technologies. Ensure that employees acquire sufficient knowledge in developing, improving, deploying or using the AI system throughout the entire life cycle.
- **Measure impact on social goals:** Identify Social Development Goals (SDGs) also known as Global Goals adopted by the United Nations [35], societal benefits/social goals, or sustainability goals where the AI system can create impact on. Ensure regular screening so that the system’s social impact aligns with ethical standards and long-term social benefits, and that the environmental issues are mitigated through sustainable practices during system operations.
- **Measure energy consumption:** Measure energy consumption and related costs during training and inference stages. Identify options to minimize the system’s carbon footprint, for example by choosing a smaller (foundational) AI model or by effective finetuning. Compare effects of model hosting on premise, on cloud and on edge (device).

4 Use Case: *StableArtists* - Generative Art in Education

4.1 Objectives of the AI system *StableArtists*

We propose an AI system, generative AI for Arts Education with the acronym *StableArtists*. The technical realization of the system is based on a custom-trained text-to-image AI model that generates images based on a given prompt or textual description. The custom-trained model is obtained through a fine-tuning process where a pre-trained base model is trained further on curated data of digitalized artwork which was previously created by students. The fine-tuned model adjusts the weights of the base model so that it can now produce images in the artistic style of the peer group of students who contributed with their artwork. The workflow for image generation by the *StableArtists* app is presented in figure 4. The main goal of this system is to build AI literacy by helping students to acquire the knowledge necessary to understand AI from a technical, ethical and user or business needs perspective as described in UNESCO’s AI competency framework for students [34].



■ **Figure 4** Steps needed to generate images by the *StableArtists* app involve collection and curation of student artworks which is used for fine-tuning a LORA model which produces AI art in the artistic style of the students.

4.2 Ethics-by-Design Approach

StableArtists allows the AI-based creation of artwork that reflects the diverse skills and styles of students from different age groups or backgrounds. The system is designed to be used in formal and non-formal educational settings. The user group consists of students under the guidance of a teacher or instructor. The development of the StableArtists system is motivated by an educational objective. Students shall learn to identify biases, acquire knowledge about ethical AI practices, and eventually become responsible citizens and remain independent actors in an increasingly AI-driven society. StableArtists provides the first use case to test the operability of the AI ethics assessment blueprint. We use the outcomes of the assessment for an ethics-by-design approach in the specification, technical realization of the diversity-sensitive AI system and its intended use. The following section is based on the results of the selected measures (c.f. appendix B) from the UNESCO ethical impact assessment [33] following the processes of the AI ethics blueprint.

4.3 Acceptability of the fairness-performance equilibrium

StableArtists has the dilemma to maximize two antagonistic metrics of the underlying AI model which are fairness and performance. Fairness is measured with respect to the representation of students who contribute artwork to the training data. Students are characterized by a set of features such as age, gender, ethnicity. The performance or quality of the model is measured by the mean esthetic value of the artwork composing the training data. For finetuned image-generation models we can fairly assume that the quality of the output is representative to the quality of the input training data. The quality, or synonymously the esthetic value, will be established by grading individual student's contributions by (i) grading by the teacher, (ii) consensus decisions by the students, or (iii) by a multi-modal AI model acting as a “judge”. The students can vote on their preferred evaluation method. The finetuned model has the task to produce images of higher quality (mean grade) than a baseline model where all data would be included in the finetuning process. To fulfill this objective, artwork of lower grades must be removed from the training data which introduces selection bias. This exclusion bias will in turn lower the representativeness of the model with respect to students. The stakeholders of the system must agree on a bias mitigation strategy to select those images which will improve the esthetic value and still balance the representation of diverse student characteristics in the training data. Different bias mitigation strategies can be mapped by a materiality matrix assessment of the AI model as shown in figure 5. Students will understand how (cultural) bias may be inherent to generative AI systems as output bias is related to bias in the seen training data. StableArtists serves as a practical example through which students will gain insights into the ethical implications of AI technologies and developing a more responsible approach to their use.

Esthetic value	High	Accept	Accept	Accept
		Reject	Accept	Accept
	Low	Reject	Reject	Accept
Baseline value				Baseline
		Low	Medium	High
		Representativeness		
				Everyone

Figure 5 AI model materiality matrix assessment. Esthetic value measures the appreciation of art. The Esthetic value of the training data correlates to the generated output images of the finetuned model. Representativeness is the measure of selection bias for excluded images in the training dataset to achieve a higher esthetic value.

5 Conclusions

The paper proposes a framework for the ethics assessment along the AI lifecycle divided in phases and processes. This blueprint is based on the concept of adaptability; the framework is agnostic to specific ethics guidelines, regulatory approaches, industry sectors, business models, and technologies. It allows to choose use-case specific measures from the selected ethics framework (e.g. UNESCO) and to prioritize the most relevant ethics KPIs from the outcomes catalog. Conducting an AI ethics assessment according to the blueprint is not merely a compliance criterion; but adds value to the overall AI system by enhancing user adoption and trustworthiness, towards achieving Trusted AI. Trusted AI in practice requires two components, first an enforceable component to achieve compliance with regulatory standards on AI quality, and second an voluntarily component built on an AI assessment blueprint for ethics-by-design approach with selectable an adaptable AI Ethics KPIs.

5.1 Recommendations further research

While the Blueprint provides a promising foundation for AI ethics assessment, further research is needed to continue refining the framework. We propose three prospective directions:

1. Develop a process for applying the blueprint in ethical assessment of potential transitions of AI applications between different risk categories with respect to the classification defined by the EU AI Act.
2. Test the adaptability of the AI ethics assessment framework through use cases in: (i) different geographical zones with different interpretability of the ethics principles, and in (ii) sensitive areas like healthcare, recruiting, AI at the workplace, or collaborative AI systems.
3. Identify generalisation aspects of AI Ethics Assessment across different application field sectors, with respect to the harmonization of outcomes, and guiding the standardisation of AI Ethics Assessment.

References

- 1 African Union. *Continental Artificial Intelligence Strategy. Harnessing AI for Africa's Development and Prosperity*. African Union, July 2024.
- 2 BCLP. US state-by-state AI Legislation snapshot. *bclplaw.com*, 2024.
- 3 Anu Bradford. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press New York, 1 edition, February 2020.
- 4 Anu Bradford. *Digital Empires: The Global Battle to Regulate Technology*. Oxford University Press, Oxford, New York, September 2023.
- 5 Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. URL: <http://proceedings.mlr.press/v81/buolamwini18a.html>.
- 6 CIFAR. Pan-Canadian Artificial Intelligence Strategy. *cifar.ca*, 2017.
- 7 Nicholas Kluge Corrêa, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, Edmund Terem, and Nythamar De Oliveira. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10):100857, October 2023. doi:10.1016/j.patter.2023.100857.
- 8 Embassy of the People's Republic of China in Grenada. Global AI Governance Initiative. *gd.china-embassy.gov.cn*, October 2023.
- 9 European Commission. Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI*. Publications Office, LU, 2019.
- 10 European Commission. Directorate-General for Communications Networks, Content and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment*. Publications Office, LU, 2020.
- 11 European Parliament and European Council. AI Act, Regulation 2024/1689. *Official Journal of the European Union*, June 2024.
- 12 Luciano Floridi. Soft Ethics and the Governance of the Digital. *Philosophy & Technology*, 31(1):1–8, March 2018. doi:10.1007/s13347-018-0303-9.
- 13 Government of Canada. Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. *justice.gc.ca*, November 2022.
- 14 Government of Canada. The Artificial Intelligence and Data Act (AIDA). *ised-isde.canada.ca*, 2023.

- 15 Government of Nigeria. National Artificial Intelligence Strategy. *ncair.nitda.gov.ng*, August 2024.
- 16 Ministry of Law and Justice. The Digital Personal Data Protection Act. *meity.gov.in*, August 2023.
- 17 Monetary Authority Singapore. Assessment Methodologies for Responsible Use of AI by Financial Institutions. *mas.gov.sg*, February 2022.
- 18 NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology, Gaithersburg, MD, January 2023. doi:10.6028/nist.ai.100-1.
- 19 NITI Aayog. National Strategy for AI #AIForAll. *niti.gov.in*, 2018.
- 20 Ricardo Ortega-Bolaños, Joshua Bernal-Salcedo, Mariana Germán Ortiz, Julian Galeano Sarmiento, Gonzalo A. Ruz, and Reinel Tabares-Soto. Applying the ethics of AI: A systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, 57(5):110, April 2024. doi:10.1007/s10462-024-10740-3.
- 21 Personal Data Protection Commission Singapore and Infocomm Media Development Authority. Model AI Governance Framework (2nd edition). *pdpc.gov.sg*, January 2020.
- 22 PRC Cyberspace Administration. Interim Measures for the Management of Generative Artificial Intelligence Services (translated). *cac.gov.cn*, July 2023.
- 23 PRC Ministry of Science and Technology. Ethical Norms for New Generation Artificial (translated). *most.gov.cn*, October 2021.
- 24 Republic of South Africa. National Artificial Intelligence Policy Framework. *dcdt.gov.za*, August 2024.
- 25 Reuters. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *reuters.com*, October 2018.
- 26 The Hollywood Reporter. Scarlett Johansson's AI Legal Threat Sets Stage for Actors' Battle With Tech Giants. *hollywoodreporter.com*, May 2024.
- 27 The Indian Express. His 'jhakaas': HC issues order against misuse of Anil Kapoor's persona. *indianexpress.com*, May 2024.
- 28 The White House. Blueprint for an AI Bill of Rights. *whitehouse.gov*, October 2022.
- 29 The White House. Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. *whitehouse.gov*, July 2023.
- 30 The White House. Delivering on the Promise of AI to Improve Health Outcomes. *whitehouse.gov*, December 2023.
- 31 The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *whitehouse.gov*, October 2023.
- 32 UNESCO. Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO, 2022. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- 33 UNESCO. Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO, 2023. doi:10.54678/YTSA7796.
- 34 UNESCO. AI competency framework for students. Technical report, UNESCO, 2024.
- 35 United Nations. Sustainable Development Goals. *sdgs.un.org*, 2024.
- 36 United Nations. AI Advisory Body. Governing AI for Humanity. Technical report, United Nations, September 2024.
- 37 Angela Huyue Zhang. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law (forthcoming)*, January 2024. doi:10.2139/ssrn.4708676.

A Appendix 1: UNESCO Recommendation on Ethics of AI

UNESCO produced the first global standard on AI ethics – the Recommendation on the Ethics of Artificial Intelligence [8]. This framework was adopted by all 193 Member States in 2021. The Recommendation is centered around 10 principles:

I. Proportionality and Do No Harm. through which AI must be used only to achieve its legitimate purposes. It must not cause harm, discriminate, or manipulate. Risk assessments must ensure AI goals are appropriate, balanced, respect human rights, and are scientifically reliable.

II. Safety and Security. where AI actors should prevent and address unwanted harms (safety risks) and vulnerabilities to attacks (security risks).

III. Fairness and Non-Discrimination. by which AI actors must ensure fairness, inclusivity, and accessibility, address biases and digital divides. Member States should promote equity, and advanced countries should support less advanced ones. Measures for discrimination must be available.

IV. The Sustainability. principle states that AI technologies can either support or hinder sustainability goals, depending on their application. A continuous assessment of their human, social, cultural, economic, and environmental impact is required to align the system with the Sustainable Development Goals (SDGs).

V. Right to Privacy, and Data Protection. recommends that privacy (imperative for human dignity and autonomy) is protected throughout the AI lifecycle. Data handling must align with international and local laws, and strong data protection frameworks should be established considering societal and ethical aspects.

VI. Human Oversight and Determination. by which member states must ensure that ethical and legal responsibility for AI systems can always be attributed to individuals or entities. Human oversight should include both individual and public oversight. Ultimate responsibility and accountability are always ascribed to humans.

VII. Transparency and Explainability. support accountability, help individuals understand AI decisions, and promote democratic oversight. UNESCO recommends that the level of transparency and explainability should be appropriate to the context of use, as there may be tensions between these two and other principles such as privacy, safety and security.

VIII. Responsibility and Accountability. principle recommends developing AI systems that are auditable and traceable. Oversight, impact assessments, audits, and whistle-blower measures are needed to avoid conflicts with human rights and environmental standards.

IX. Awareness and Literacy. states that the public awareness of AI and data must be increased through open education, civic engagement, civil society actions, academia and the private sector involvement, etc. AI education needs to address its impact on human rights, freedoms, and the environment.

X. Multi-Stakeholder and Adaptive Governance and Collaboration. calls on data use to respect international law and national sovereignty, allowing states to regulate data within their territories and ensure data protection while upholding privacy rights. Stakeholder participation is needed to achieve inclusive AI governance, involving governments, organizations, the technical community, civil society, academia, media, policymakers, and others. Participation from marginalized groups and Indigenous Peoples is contributing to sustainable development and effective AI governance.

In alignment with the above-mentioned principles, following the need for an AI impact assessment, UNESCO developed a methodology for ethical impact assessment of AI systems in 2023. The methodology was published in the document *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of AI* [33]. The goal of the assessment is to ensure alignment of AI system with values and principles recognized by UNESCO in the Recommendation. However, there is still a step to go from endorsement of a recommendation by governments to an actual implementation of the ethical impact assessment by AI producers in practice.

B Appendix 2: Use Case *StableArtists*

Fig. 6 shows the selected UNESCO framework measures mapped to the blueprint for AI ethics assessment which we conducted on the generative AI use case *StableArtists*. A description of the measures is given below.

Questions for phase 1

- *Q-111*: Please provide an initial description of the AI system you intend to design, develop or deploy:
- *Q-112*: Please describe the aim or objective of this system. If the aim is to address a specific problem, please specify the problem you are trying to solve. Please also specify how this system may fit within broader schemes of work:
- *Q-1141*: Who will the users who interact with your system be (include their level of competency)?
- *Q-62214*: Have you developed a process to document how data quality issues can be resolved during the design process?

Questions for phase 2

- *Q-6232*: How has the principle of fairness been approached from a technical perspective? For example, are you able to specify what the technical notion of fairness is that the AI system is calibrated for? (e.g., individual fairness, demographic parity, equal opportunity, etc.)
- *Q-4245*: Which activities will help your team to identify potential impacts and ensure they are mitigated?

Questions for phase 3

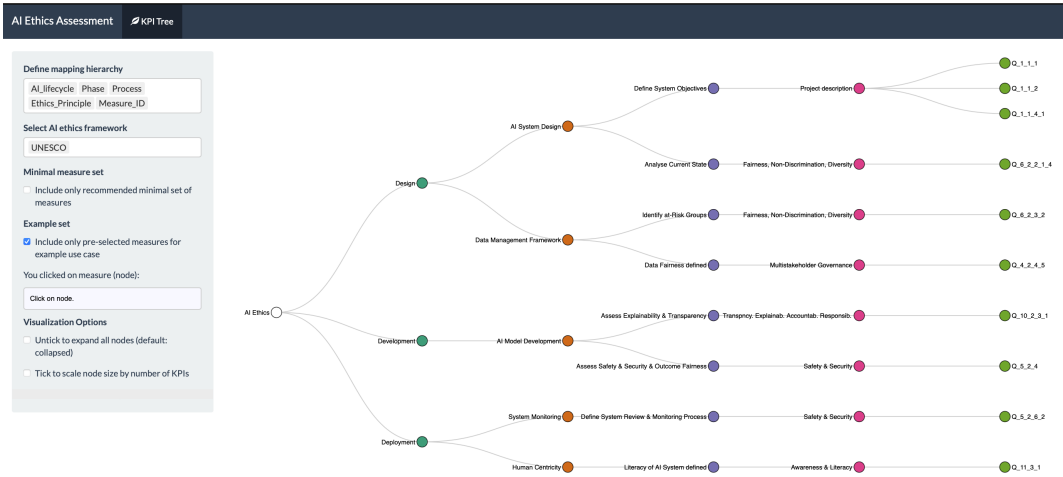
- *Q-10231*: Is the algorithm, including its inner-working logic, open to the public or any oversight authority? Is the code of the AI system in an open-source format?
- *Q-524*: If the training data or data being processed by the AI system were poisoned or corrupted, or if your system was manipulated, how would you know?

Questions for phase 4

- *Q-5262*: How often will the AI system be tested in the future and which components will be tested?

Questions for phase 5

- *Q-1131*: What are the prospective positive impacts of the system on AI awareness and literacy? How, if at all, could the deployment of this system increase awareness surrounding AI? Are there any other ways in which this system could increase awareness and literacy?



■ **Figure 6** Selected framework measures for the use case *StableArtists*.

AI Readiness of Standards: Bridging Traditional Norms with Modern Technologies

Adrian Seeliger 

Deutsches Institut für Normung e.V. (DIN), Berlin, Germany

Abstract

In an era where artificial intelligence (AI) is spreading throughout most industries, it is imperative to understand how existing regulatory frameworks, particularly technical standards, can adapt to accommodate AI technologies. This paper presents findings of an interdisciplinary research & development project aimed at evaluating the AI readiness of the German national body of standards, encompassing approximately 30,000 DIN, DIN EN, and DIN EN ISO documents. Utilizing a hybrid approach that combines human expertise with machine-assisted processes, we sought to determine whether these standards meet the conditions required for secure and purpose-specific AI implementation.

Our research focused on defining AI readiness, operationalizing this concept, and evaluating the extent to which existing standards meet these criteria. AI readiness refers to whether a standard complies with the conditions necessary for ensuring that an AI system operates securely and as intended. To operationalize AI readiness, we developed explicit criteria encompassing AI-specific requirements and the contextual application of these standards. A dual approach involving thorough human analyses and the use of software automation was employed. Human experts annotated standardization documents to create high-quality training data, while machine learning methodologies were utilized to develop AI models capable of classifying the AI readiness of these documents.

Three different software tools were developed, to provide a proof-of-concept for a more scalable and efficient review of the 30,000 standards. Despite certain technical and organizational challenges, the integration of both human insight and machine-led processes provided valuable and actionable results and insights for further development.

Key findings address the exact choice of words and graphical representation in standardization documents, normative references, categorization of standardization documents, as well as suggestions for concrete document adaptations.

The results underscore the importance of an interdisciplinary approach, combining domain-specific knowledge and advanced AI capabilities, to future-proof the intricate regulatory frameworks that underpin our industries and society.

2012 ACM Subject Classification Proper nouns: People, technologies and companies → International Organization for Standardization; Computing methodologies → Artificial intelligence; Proper nouns: People, technologies and companies → European Telecommunications Standards Institute

Keywords and phrases Standardization, Norms and Standards, AI Readiness, Artificial Intelligence, Knowledge Automation

Digital Object Identifier 10.4230/OASICS.SAIA.2024.8

Category Practitioner Track

Acknowledgements We extend our gratitude to the Bundesministerium für Wirtschaft und Klimaschutz (BWMK). Our sincere thanks for their invaluable support and guidance throughout this project to the Fraunhofer IAIS Team “AI Safeguarding and Certification” and Team “Natural Language Understanding” for their expertise and collaboration. We also wish to thank Fraunhofer IKS, IEM, HHI, MEWIS, and INT for their significant contributions without which this project would not have been possible.



© Adrian Seeliger;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 8; pp. 8:1–8:6

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

AI as an emerging technology will impact most areas of industry and society. The project „AI Readiness of Standards“ was set to find out what this means for the German national body of standards, which consists of roughly 30,000 DIN, DIN EN and DIN EN ISO documents.

The approach was a mixture of individual human assessment and a machine-assisted process.

During the 2.5 years of the project, answers to the following research questions were sought:

- How can a complex, 100-year-old regulatory framework be adapted for AI technologies?
- Is a comprehensive review of all 30,000 documents for AI readiness necessary or possible?
- How crucial is the application context in adapting standards for AI?
- How might software technologies help with this?

This paper gives an overview of the findings.

Section 2 gives an overview of definitions and previous research, setting out the fundamental principles for this project.

These principles were applied by human experts to a large number of existing standards documents, resulting in section 3, which provides a high-level overview of the challenges and results, including a brief discussion.

To be able to apply the principles to the entirety of the body of standards in an economically feasible way, automation technology has to be used. Section 4 describes the development and use of three primary software tools.

Finally, section 5 concludes the paper with the projects' achievements and possible future challenges.

2 Methodology: Defining and Operationalizing AI Readiness

What actually is AI readiness? In current literature, there are several understandings of AI readiness, mostly revolving around organization theory or innovation adoption frameworks. AI readiness generally refers to an organization's capacity to deploy and utilize AI technologies effectively. The concept encompasses various dimensions including technological infrastructure, organizational preparedness, and the external environment.

For instance, one study describes AI readiness as part of the broader process of AI adoption, emphasizing the need for ongoing assessment and development rather than a single preliminary evaluation [4]. It posits that AI readiness and adoption are highly interdependent and must be integrated throughout the entire adoption process to ensure successful implementation. Another work offers an AI readiness framework which assesses an organization's capabilities in four key dimensions: technologies, activities, boundaries, and goals, providing a pragmatic tool for facilitating digital transformation within organizations [3].

[2] highlight technological, organizational, and environmental factors, while [1] define AI readiness as an organization's capacity to implement and utilize AI using cultural aspects and necessary resources, emphasizing the importance of top management support, resource availability, and organizational infrastructure.

Regarding standardization there was no known precedence as to how a standards document that is AI ready should look like, so we defined the term ourselves as follows (summarized):

The AI readiness of a standard refers to whether it meets conditions that ensure an AI system compliant with this standard is secure and remains so according to its intended purpose. A standard is considered AI ready if it is specific enough to cover the use of AI or AI-specific requirements and measures, while being unambiguous for users.

For more information see [5].

The initial definition and the evaluation method were developed through lengthy discussions and a wide stakeholder participation process, engaging AI experts, DIN standards committees, business associations as well as the public. This generally worked well. But already at this stage there were often more questions than answers.

While later applying the evaluation method, feedback loops provided input for 20 subsequent evolutions. Validation and testing in real conditions were identified as key for success, but not trivial.

Different industry experts for the application of standardization documents in their respective field proved rare, leading to a higher level of human resources needed to generate expert knowledge.

Additionally, the lack of legal expertise in the project led to the deferral of that aspect, addressing it in a dedicated step within the evaluation method.

3 Human Expert AI Readiness Assessment

The high complexity of the task “detailed human expert analysis of approximately 1100 standardization documents” made it challenging to ensure consistency in evaluation across different experts. The evaluation of normative documents was heavily influenced by the evaluators’ background knowledge, personal biases, and their interpretation of the content e.g., resulting in a different understanding of concepts such as AI, practicability of AI applications in a specific domain. Efforts such as verifying the plausibility of classifications and resolving conflicts through expert reviews were taken. This process required an additional layer of review and revision, adding complexity to the workflow. A certain level of bias and interpretive flexibility remained.

3.1 General AI Potentials

The integration of artificial intelligence (AI) in various fields holds significant potential for improving efficiency, accuracy, and consistency. To optimize these benefits, it is crucial to avoid subjective formulations of human cognitive abilities by providing explicit and objective descriptions. This ensures a clear understanding and application of AI capabilities without human biases.

Standardizing tables and figures is essential for unifying diverse representations of knowledge. By assigning clear and distinguishable names to different formats (e.g., differentiating between a sketch and a technical drawing, or a tolerance table and a test procedure table), computers can more effectively process and differentiate this information.

Identifying the type of document (e.g., product standards, process standards) beforehand, rather than relying on text searches within the document, can significantly streamline information retrieval and processing. Additionally, simplifying cross-references to other normative documents in a digitally capturable format, including reference cascades, would greatly enhance accessibility and usability.

AI risks and quality requirements should be marked by references to horizontal standards. This not only ensures compliance with safety standards but also provides a framework for assessing and managing potential AI-related risks.

3.2 AI Potentials in Mechanical Engineering

Throughout the whole project a variety of computer vision applications have been found to be the main driver for AI relevance of standardization documents. Developing horizontal standards for visual inspections would standardize this critical, frequently referenced process, ensuring uniformity and reliability.

Similarly, creating standards for document management, automated document inspections, and the generation of test reports would streamline these activities, enhancing efficiency and accuracy.

To facilitate the safe use of AI, especially in safety-critical applications, standards should regulate language variability and restriction, making programming languages more predictable and secure. ISO/IEC TR 5469 could serve as a sector-specific AI horizontal standards, potentially including human or programmatic control instances to ensure oversight and accountability.

Further, there is a need for new standards governing AI's use in control systems. This includes clarifying the classification and application of AI within Safety Integrity Level (SIL) and Performance Level (PL) frameworks. Adapting existing standards such as ISO 13849-1 and -2, and mapping AI performance to current SIL/PL levels, will help maintain functional safety standards. The use of AI systems as redundancies can enhance safety by providing backup options that meet or exceed the reliability of human-operated systems.

3.3 AI Potentials in Healthcare

The medical field would benefit from sector-specific AI horizontal standards for various applications. These include antibiogram analysis, the generation of test reports, and the linking and evaluation of diverse sensory data. For example, in microbiology, AI could aid in the classification of bacteria and the measurement of growth rates (cell colony count). In radiology, AI can enhance the identification of human cell types, such as distinguishing between tumor cells and tissue cells.

Additionally, existing standards, like the risk management protocol for medical devices (DIN EN ISO 14971), should be reviewed and potentially enhanced to align with the advancements in AI technology. This will ensure that AI applications in healthcare maintain the rigorous safety and quality standards required in this critical sector.

3.4 AI Potentials in Automotive

In the automotive sector, implementing clustering and introducing specific categories or adapting an A/B/C type system, as seen in the functional machine safety domain, would streamline various functions. These categories should cover interface descriptions, specifications, architecture descriptions, test standards, and procedure descriptions. Such a structured approach would enhance the clarity and efficiency of AI applications in automotive engineering.

4 Automation Attempts: Machine Assisted AI Readiness Assessment

The national body of standards has traditionally been dominated by technical drawings, specifications, and historically developed document families. However, in contemporary applications, there has been a significant shift towards digitization, with increasing demands for “smart” standards, automated integration of standard contents, and even automated conformity checks.

Given the vast number of approximately 30 000 standards, a meticulous manual review is deemed impractical. Consequently, this project explored the utilization of software for the large-scale and scalable analysis of these standards.

Specifically, the research focused on the conceptualization, development, and investigation of three primary tools: a tool for semantic similarity search, a tool for linking standardization documents to a knowledge database, and an AI tool designed for the automated classification of standards into AI readiness classes.

4.1 Semantic Similarity Search

The Semantic Similarity Search tool was developed to find text similarities. It leverages the representational capabilities of transformer encoder models, such as BERT, to generate meaningful vector representations of documents. These vector representations capture the semantic essence of the input text, facilitating the identification of semantically similar paragraphs, sentences, or whole documents. The goal of this tool was to ease (manual) classification of documents by finding similar documents to already classified ones. Due to delays in implementation, Semantic Similarity Search was not extensively used in the project.

4.2 Knowledge Database Tool

The Knowledge Database Tool was created to link the entire body of standards to the existing open source knowledge database OpenAlex. Comparing the standards' title and short description to the structured data of 200 million publications enabled calculating a specific AI-distance-score for each standards document. A low AI-distance score could signal a close relation of that document to AI technologies and thus a higher level of AI relevance with possible implications for AI readiness. Furthermore, the citation network of standards citing other standards was analyzed, resulting in similar AI distance score of standardization documents with mutual citations.

4.3 AI Tool

The AI tool is designed to automate the classification of AI readiness of standardization documents. It consists of three components:

The annotation component enabled the labelling of standardization documents by human experts, producing the training data necessary for developing the classification mechanisms of the AI tool.

The AI model is the core classification module, developed using advanced machine learning methodologies and utilizing the curated training data of the annotation component to categorize the AI readiness of standardization documents.

The User Interface (UI) offers an intuitive platform to showcase the AI model's results.

Technical & Organizational Challenges

- Issues such as the inability to annotate graphical elements (like graphs, schematics, and tables) due to technical restrictions in the labelling tool posed significant challenges.
- The AI model could not be fully trained with the evaluating experts external knowledge. This includes information from referenced standards, application context and implicit knowledge.

- To ensure copyright protection and information security, measures such as the use of external servers, specialized VPN access were implemented. These measures, combined with the collaboration of multiple parties, extended the initialization phase and complicated the resolution of issues like data exchange and server failures, making the process more laborious and time-consuming.
- Watermarks in PDF documents caused difficulties as they could be on a different layer than the normal text. While labelling, the tool often couldn't distinguish between these layers, leading to the unintended annotation of entire pages. Watermarks were required for information security reasons.
- The labelling tool used was robust but had high memory requirements which sometimes led to technical issues such as latency and synchronization problems. The workflow had to be adapted to accommodate the tool's limitations and the extensive memory needed led to occasional disruptions.
- Users reported that selecting a label and marking the corresponding text passage appeared smooth initially. However, changing the label selection to create a new annotation often led to a lag. This delay sometimes resulted in incorrect label selection and subsequent inaccurate annotations.
- The annotation component struggled to handle text that spanned multiple pages.
- The PDF import feature occasionally required manual parsing for correct uploading. If there were problems with PDF uploads, TXT formats were used instead. This involved extracting text from PDF files using either an XML parser or a PDF parser, with a preference for XML due to its structured nature and higher accuracy.

5 Conclusion

The project aimed to develop a proof-of-concept and provide initial concrete indications and points of connection for AI in established standards. Both aspects were clearly answered positively and supported with examples and results.


However, as is often the case, complex issues require complex solutions. A systematic integration of AI into the standards framework requires deepening the developed content and building competencies. Software and data-driven approaches can help, but they also come with non-negligible additional organizational and technical efforts. We are curious to see how the advancing digitalization of our society will impact standardization and look forward to future challenges.

References

- 1 Wajid Ali and Abdul Zahid Khan. Factors influencing readiness for artificial intelligence: a systematic literature review. *Data Science and Management*, 2024.
- 2 Sulaiman Alsheibani, Yen Cheung, and Chris Messom. Artificial intelligence adoption: Ai-readiness at firm-level. In *Pacific Asia Conference on Information Systems 2018*, page 37. Association for Information Systems, 2018. URL: <https://aisel.aisnet.org/pacis2018/37>.
- 3 Jonny Holmström. From ai to digital transformation: The ai readiness framework. *Business Horizons*, 65(3):329–339, 2022. doi:10.1016/j.bushor.2021.03.006.
- 4 Jan Jöhnk, Malte Weißert, and Katrin Wyrski. Ready or not, ai comes—an interview study of organizational ai readiness factors. *Business & Information Systems Engineering*, 63(1):5–20, 2021. doi:10.1007/S12599-020-00676-7.
- 5 Kostina Prifti, Esra Demir, Julia Krämer, Klaus Heine, and Evert Stamhuis, editors. *Digital Governance: Confronting the Challenges Posed by Artificial Intelligence*. Information Technology and Law Series. T.M.C. Asser Press The Hague, 1 edition, 2024. Hardcover due: 08 January 2025, Softcover due: 08 January 2026, eBook due: 08 January 2025.

Introducing an AI Governance Framework in Financial Organizations

Best Practices in Implementing the EU AI Act

Sergio Genovesi 

SKAD AG, Frankfurt am Main, Germany

Abstract

To address the challenges of AI regulation and the EU AI Act's requirements for financial organizations, we introduce an agile governance framework. This approach leverages existing organizational processes and governance structures, integrating AI-specific compliance measures without creating isolated processes and systems. This framework combines immediate measures to address urgent AI compliance cases with the development of a broader AI governance. It starts with an assessment of requirements and risks, followed by a gap analysis; after that, appropriate measures are defined and prioritized for organization-wide execution. The implementation process includes continuous monitoring, adjustments, and stakeholder feedback, facilitating adaptability to evolving AI standards. This procedure guarantees not only adherence to current regulations but also positions organizations to be well-equipped for prospective regulatory shifts and advancements in AI applications.

2012 ACM Subject Classification General and reference → Empirical studies

Keywords and phrases AI Governance, EU AI Act, Gap Analysis, Risk Management, AI Risk Assessment

Digital Object Identifier 10.4230/OASICS.SAIA.2024.9

Category Practitioner Track

Acknowledgements Many thanks to Dennis Kautz and Kim Strunk for their valuable feedback and insights. Thanks to Felix Broßman, Daniel Schulz und Helge Krill for their trust and support.

1 AI Regulation and the Financial Sector

Artificial Intelligence (AI) is transforming the financial sector by powering advisory services, enhancing risk management, and improving compliance and fraud detection. AI drives innovation through automated data analysis and personalized marketing and sales strategies while boosting operational efficiency. However, AI's rapid integration into business processes brings substantial challenges regarding regulatory complexities and potential legal issues [7]. Implementing AI can create unanticipated risks, necessitating tailored approaches for safe operation. Moreover, AI incidents can impact public perception and pose reputational concerns. Traditional risk management approaches lack the required technical depth and fall short in addressing the various AI implications, highlighting the need for specialized AI governance practices.

In the European context, to use and deploy AI in the financial sector, it is necessary to navigate a complex regulatory environment – recently intensified by the addition of the EU AI Act [12]. The EU AI Act joins a suite of important existing regulations, such as the General Data Protection Regulation (GDPR) [8], the Digital Markets Act [9], the Data Act [11] and the Data Governance Act [10], as well as sector-specific rules and standards such as the EBA Guidelines [2] or, in Germany, the German Banking Act [1] and the BaFin minimal requirement for risk management (MaRisk) [6], collectively shaping the responsible deployment and management of the digital infrastructure and services in the financial sector.



© Sergio Genovesi;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 9; pp. 9:1–9:7

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The EU AI Act adopts a risk- and technology-based regulatory approach, introducing different risk categories for AI systems, ranging from unacceptable, high, or systemic risk to limited risk. High-risk applications, used in sensitive fields like credit scoring or recruiting, which could have adverse effects on individuals and society, will be subject to transparency, data governance, and risk and quality management requirements – among others. Sanctions for non-compliant organizations are strict, with fines reaching €35 million or up to 7% of global annual turnover, emphasizing the critical need for financial organizations to enhance their governance structures and implement rigorous monitoring and audit procedures.

These requirements build upon those for data technology and IT assets introduced by the aforementioned regulations, adding AI-specific obligations. For instance, Article 10 of the EU AI Act, which regulates Data and Data Governance, aligns with GDPR provisions regarding the processing of personal data while introducing data governance practices aimed at training, validation, and testing datasets for AI systems.

In the context of establishing a systematic risk management framework, the MaRisk requirements already set an important benchmark concerning risk assessment and management for IT systems, including both technical and organizational measures to ensure the integrity, availability, authenticity, and confidentiality of the data. While addressing new AI-specific threats is a defining feature of the AI Act, the overlaps concerning “classic” IT risks affecting AI systems as well as the design of risk management frameworks are evident. The EU AI Act itself encourages the integration of AI-specific risk management measures within existing frameworks (see, for example, Article 9, paragraph 10). As we will see more specifically in the next section, these overlaps allow organizations to include new AI specific compliance measures in their existing compliance frameworks.

2 Implementing an AI Governance Strategy in Financial Organization

2.1 Assigning responsibilities within existing functions

Since the AI governance requirements of the EU AI Act intersect with several operational fields that are already subject to a variety of regulations, a strategic incorporation of new organizational measures across different departments is necessary.

To individuate the most suitable business processes and departments for the implementation of AI requirements in specific operational fields, it is possible to refer to the three lines of defence (LoD) model – a widespread risk governance framework in financial organizations [3]. According to Engels et al., “[t]he 3LoD model caters for coordination of control responsibilities among various stakeholders [...]. This is achieved by allocating and delineating distinct roles and mandates to business and operational functions in the 1st LoD, internal control and standard setting functions in the 2nd LoD as well as internal audit in the 3rd LoD”. [3, p. 97]. Focusing specifically on the internal controls established by different functions in the 2nd LoD, it is possible to highlight possible overlaps and synergies between existing controls and new measures. This analysis facilitates a thematic clustering and helps delineate precise task, enabling a clearer division of responsibilities for assigning working packages necessary for implementing specific requirements.

The following list includes operational fields within the 2LoD that according to our analysis are affected by the EU AI Act, and briefly describes to what extent:

- **Information Security / Risk Management** faces significant implications, including the need for AI-specific asset categorization and vulnerability assessments, as well as the management of unique AI risks and a reevaluation of compliance measures.

- **Operational Risk Management** must calculate and allocate sufficient resources for potential AI-related damages and consider insurance options specifically for AI risks, acknowledging that AI can introduce new variables and uncertainties into the operational risk framework.
- **Data Governance** is a critically impacted area, necessitating new measures addressing the accuracy of output, possible biases in training data, and the overarching management of data quality, among other things.
- **Outsourcing and Vendor Management** is affected not only by the necessity for robust contractual clauses that cover AI specifics but also by the heightened attention to supply chain risks and the requirement for continuous risk analysis and monitoring of service level agreements (SLAs) in relation to AI suppliers and vendors.
- **Compliance** functions will need to adjust existing practices to ensure adherence to new legal AI regulations and requirements, and stay up to date with the latest standards to ensure AI trustworthiness. Especially concerning Generative AI use cases, it will be important to understand the implications for copyright and trademark laws concerning training data and generated content.
- **Data Protection** teams are responsible for ensuring the proper management of personal data within AI systems, overseeing transnational data processing activities, and maintaining the security of sensitive data categories.

Effective collaboration between the aforementioned functions and IT and other specialized departments is crucial for the successful adherence to AI regulatory requirements. Moreover, communication among affected stakeholders to coordinate implementation measures in case of partial overlap is key to optimizing the division of tasks and ensuring the success of the new governance strategy. This cross-departmental cooperation is likely to feature the utilization of decentralized risk management modules as integral components of the frontline risk management strategy. To ensure that all relevant risks are addressed, the organization's management board may decide to appoint a new role to supervise the progress of work. This role can be integrated into existing departments and added to the responsibilities of current positions.

2.2 Agile Implementation Methodology

Recognizing the overlap with other ongoing risk governance processes, the proposed approach does not create additional structures dedicated to AI governance and EU AI Act compliance, but integrates and expands upon existing frameworks. This integration ensures that the measures taken are rooted in the existing operational structure, promoting a seamless transition to an AI-ready compliance and risk management strategy.

Adopting this strategy initiates with a comprehensive briefing and planning phase. This step involves a thorough scoping of the organization's needs and strategic goals, defining the level of ambition, and identifying key stakeholders. Informed by this insight, project planning and governance are established and consolidated in a collaborative kick-off with all stakeholders.

This is followed by an assessment phase dedicated to understanding the applicable regulatory environment and the identification of AI-specific risks. At this stage, the status quo with regard to AI governance is investigated as well. These assessments lay the groundwork for the gap analysis phase, where existing governance is evaluated against new regulatory standards, allowing for the precise definition and prioritization of necessary measures that span across all levels of the organization.

Once defined, the new measures are systematically implemented. During the implementation phase, regulatory developments are closely monitored to ensure ongoing alignment and fast adjustment to new standards and requirements. Finally, the process naturally progresses to an optimization phase, which aims at bringing iterative enhancements to the organization (including communication strategies and customized training). This phase allows for the establishment of a Continuous Improvement Process (CIP) for the ongoing development of AI governance, rounding off with the comprehensive completion of the project and assimilation of stakeholders' feedback.

This strategy adopts an agile framework, incorporating both bottom-up and top-down procedures. Bottom-up workflows address immediate compliance needs by focusing on AI use cases necessitating urgent action. Simultaneously, an overarching AI governance framework is designed based on a top-down analysis of the EU AI Act requirements and of the identified compliance gaps, pinpointing necessary enhancements to organizational processes and systems.

In applying this agile methodology, strategy developers ensure a prompt response to immediate compliance requirements while constructing a comprehensive, resilient, and sustainable AI governance infrastructure that is well-equipped for future advancements.

2.3 Examples: Best Practices and First Steps within different 2nd LoD functions

In this section, examples of first steps in implementing the EU AI Act requirements in two different 2LoD are presented. The functions considered are Information Security and Risk Management, and Outsourcing and Vendor Management.

Information Security and Risk Management

To comply with the AI Act, organizations must first assess AI-specific risks and undertake a risk classification for each AI system in use or development. Since IT risk assessments are usually performed by information security departments, the classification of AI systems and the attribution of adequate risk management controls can be integrated into the existing Information Security Management System (ISMS) and supported by the Governance, Risk, and Compliance (GRC) software already in place.

To facilitate the governance objective of identifying and classifying AI specific risks, the following first steps can be taken:

- Expand the threat catalogue in order to include potential threats and vulnerabilities associated with AI systems. In order to do so, the ENISA Threat Landscape reports published in 2020 [4] and 2023 [5] can be taken as a reference. This task involves evaluating AI specific points of vulnerability and new kind of cyberattacks targeting AI inherent weaknesses.
- Update the Configuration Management Database (CMDB) and the IT Asset Inventory to include AI-specific risk by introducing new IT asset categories that explicitly classify AI systems based on their risk levels. Considering the definitions included in Article 3, 6, 50 and 51 of the EU AI Act, the following categories should be included: AI system, high-risk AI system, general purpose AI system, general purpose AI system with systemic risk, AI system with transparency obligations. Additionally, any AI systems that fall under the forbidden practices outlined in Article 5 must be immediately discontinued.

- Compare current risk management measures with AI-specific ones to assess the necessity of adjusting existing cybersecurity protocols for mitigating AI-related threats. If current measures already fulfill the AI Act requirements (such as those concerning IT documentation and logging activities), introducing new controls should be avoided to prevent redundancies in the ISMS. Only unaddressed requirements necessitate new controls.

Once high-risk-systems are correctly flagged in the ISMS and in the GRC software, it is possible to implement adequate risk management measures to ensure the systems meet the EU AI Act requirements. To achieve this, it is necessary to focus on several key actions that integrate AI-specific controls and processes into existing information security frameworks. These actions include:

- Identify specific controls for AI systems and map these controls to relevant use cases within the organization. In alignment with the agile implementation strategy presented above, as well as with the EU AI Act provisions for risk management systems (Art. 9), the feasibility of integrating these controls into existing risk management systems should be evaluated.
- Establish a system for ongoing compliance assessment to ensure adherence to established controls. Whenever possible, automate the steps of the compliance assessment to increase efficiency.
- Establish clear decision-making processes, with a special focus on risk acceptance. This entails consulting relevant internal stakeholders throughout the AI system life-cycle to assign specific roles and responsibilities.

Outsourcing and Vendor Management

Many IT products and services used by financial organizations are not developed internally but are provided by third-party vendors. Thus, implementing the EU AI Act requirements necessitates robust due diligence and effective management of these third-party relationships to ensure compliance and mitigate risks associated with outsourcing and vendor management.

To facilitate the governance objective of third-party due diligence, the following first steps can be taken:

- Flag AI-related third-party relationships within the organization's contract database or outsourcing register. This helps in identifying all relevant stakeholders who contribute to or influence AI systems used by the organization.
- Adjust third-party assessment processes by incorporating AI-specific risk categorization and requirements. This involves refining existing risk assessment methods to address the unique risks associated with AI technologies.
- Enforcing requirements in third-party relationships by creating specific contract clauses for AI systems that cover data processing, security requirements, limitation of liabilities, and other critical aspects as mandated by the EU AI Act.

Regarding the governance objective of managing third-party AI systems, the following actions should be taken:

- Create the necessary information basis by establishing links between relevant systems, such as the ISMS and the contract database or outsourcing register, to ensure information is consistently updated and accessible to those managing third-party relationships.
- Identify, assess, and manage third-party AI-risks on an ongoing basis. This involves regular third-party risk assessments, followed by appropriate risk management measures.
- Conduct routine audits and reporting activities concerning relevant contractual partners to check for non-compliance with the EU AI Act.

3 Conclusion

Adapting to the EU AI Act requires financial organizations to integrate AI-specific compliance measures into their existing governance frameworks. By leveraging an agile implementation strategy, organizations can address immediate regulatory requirements and build a robust AI governance structure for the long term.

The regulatory landscape for AI is expected to evolve, with new rules and challenges emerging as AI technologies advance. Financial organizations should prepare for continuous adjustments to their governance practices and anticipate further regulatory changes. By adopting a flexible and integrated approach to AI governance, they can ensure compliance with the current EU AI Act requirements while preparing for future regulatory and technological advancements.

References

- 1 Bundesregierung. Gesetz über das Kreditwesen (Kreditwesengesetz), 2023. Available online: <https://www.gesetze-im-internet.de/kreditwg/index.html> (Accessed: 2024-08-24).
- 2 European Banking Authority (EBA). Revised guidelines on outsourcing arrangements, 2019. Available online: <https://www.eba.europa.eu/activities/single-rulebook/regulatory-activities/internal-governance/guidelines-outsourcing> and PDF: <https://www.eba.europa.eu/sites/default/files/documents/10180/2551996/38c80601-f5d7-4855-8ba3-702423665479/EBA%20revised%20Guidelines%20on%20outsourcing%20arrangements.pdf> (Accessed: 2024-08-24).
- 3 Dr. Oliver Engels, Marc Peter Klein, Peter Gürtlschmidt, Dr. Georg Lienke, and Rei Tanaka. The three lines of defence model: Key success factors for effective risk management. In *Non-Financial Risk Management in the Financial Industry*, pages 71–88. Frankfurt School Verlag, 2022. Available at: https://www.frankfurt-school-verlag.de/programm/non_financial_risk_management.html. URL: https://www.frankfurt-school-verlag.de/programm/non_financial_risk_management.html.
- 4 European Union Agency for Cybersecurity (ENISA). Artificial intelligence cybersecurity challenges, 2020. Available online: <https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges> (Accessed: 2024-09-12).
- 5 European Union Agency for Cybersecurity (ENISA). Enisa threat landscape 2023, 2023. Available online: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023> (Accessed: 2024-09-12).
- 6 Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin). Rundschreiben 05/2023 (BaFin) - mindestanforderungen an das risikomanagement - marisk, 2023. Available online: https://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Rundschreiben/2023/rs_05_2023_MaRisk_BA.html (Accessed: 2024-08-24).
- 7 Dr. Jochen Papenbrock, Dr. John Ashley, Dr. Georg Lienke, Florian Seiferlein, and Norbert Gittfried. Optimising effectiveness and efficiency: Deployment of artificial intelligence in non-financial risk management. In *Non-Financial Risk Management in the Financial Industry*, pages 213–239. Frankfurt School Verlag, 2022. Available at: https://www.frankfurt-school-verlag.de/programm/non_financial_risk_management.html. URL: https://www.frankfurt-school-verlag.de/programm/non_financial_risk_management.html.
- 8 European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. Available online: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (Accessed: 2024-09-12).

- 9 European Union. Regulation (eu) 2022/1925 of the european parliament and of the council of 14 september 2022 on contestable and fair markets in the digital sector and amending directives (eu) 2019/1937 and (eu) 2020/1828 (digital markets act), 2022. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R1925>(Accessed: 2024-09-12).
- 10 European Union. Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance and amending regulation (eu) 2018/1724 (data governance act), 2022. Available online: <https://eur-lex.europa.eu/eli/reg/2022/868/oj>(Accessed: 2024-09-12).
- 11 European Union. Regulation (eu) 2023/2854 of the european parliament and of the council of 13 december 2023 on harmonised rules on fair access to and use of data and amending regulation (eu) 2017/2394 and directive (eu) 2020/1828 (data act), 2023. Available online: <https://eur-lex.europa.eu/eli/reg/2023/2854>(Accessed: 2024-09-12).
- 12 European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence, 2024. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>(Accessed: 2024-09-12).

Evaluating Dimensions of AI Transparency: A Comparative Study of Standards, Guidelines, and the EU AI Act

Sergio Genovesi ✉ 

SKAD AG, Frankfurt am Main, Germany

Martin Haimerl ✉ 

Universität Furtwangen, Germany

Iris Merget ✉

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany

Samantha Morgaine Prange ✉

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany

Otto Obert ✉

Main DigitalEthiker GmbH, Karlstadt am Main, Germany

Susanna Wolf ✉

DATEV eG, Nürnberg, Germany

Jens Ziehn ✉ 

Fraunhofer IOSB, Karlsruhe, Germany

Abstract

Transparency is considered a key property with respect to the implementation of trustworthy artificial intelligence (AI). It is also addressed in various documents concerned with the standardization and regulation of AI systems. However, this body of literature lacks a standardized, widely-accepted definition of transparency, which would be crucial for the implementation of upcoming legislation for AI like the AI Act of the European Union (EU). The main objective of this paper is to systematically analyze similarities and differences in the definitions and requirements for AI transparency. For this purpose, we define main criteria reflecting important dimensions of transparency. According to these criteria, we analyzed a set of relevant documents in AI standardization and regulation, and compared the outcomes. Almost all documents included requirements for transparency, including explainability as an associated concept. However, the details of the requirements differed considerably, e.g., regarding pieces of information to be provided, target audiences, or use cases with respect to the development of AI systems. Additionally, the definitions and requirements often remain vague. In summary, we demonstrate that there is a substantial need for clarification and standardization regarding a consistent implementation of AI transparency. The method presented in our paper can serve as a basis for future steps in the standardization of transparency requirements, in particular with respect to upcoming regulations like the European AI Act.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases AI, transparency, regulation

Digital Object Identifier 10.4230/OASICS.SAIA.2024.10

Category Academic Track

Acknowledgements We would like to thank our fellow members of the German Standardization Roadmap on Artificial Intelligence, supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) on the basis of a decision by the German Bundestag; the Foundations working group; the Ethics sub-working group; and especially the organizing committee of DIN, the German Institute for Standardization, and DKE, the German Commission for Electrical, Electronic & Information Technologies of DIN and VDE, for providing the basis for the activities of this working group and in particular this publication.



© Sergio Genovesi, Martin Haimerl, Iris Merget, Samantha Morgaine Prange, Otto Obert, Susanna Wolf, and Jens Ziehn;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 10; pp. 10:1–10:17

OpenAccess Series in Informatics



Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction and Motivation

Transparency has been identified as one of the key features for trustworthy AI by the international expert community [1, 3, 4, 5, 6, 8, 10, 16, 21]. However, the existing body of literature – ranging from academic papers, policy documents, recommendations, to regulations, and standards – lacks a standardized, widely-accepted definition of transparency. The documents provide varying interpretations, focusing on different dimensions of transparency such as traceability of data origin, explainability of algorithmic decisions, disclosure of the system’s abilities and limitations, and assuring user awareness about them interacting with a machine, among others. This fragmentation of interpretations leads to differing requirements for achieving transparency in AI systems, posing a significant challenge for standardization and compliance assurance with respect to AI quality.

This paper is a pilot study aiming at establishing a methodology to highlight both discrepancies and commonalities across key documents, providing a tool for policy makers to identify relevant features that need to be addressed to produce effective standards for AI transparency within a specific regulatory framework. For our analysis, we focused on selected standards and guidelines of high relevance within the European framework published by prominent German, European and international organizations during the drafting phase of the European AI Act. The AI Act itself was also included as an important reference.

By considering further documents representing other perspectives and fields of interests, our methodology has the potential to scale up to other theoretical, practical, and regulatory frameworks with differing geographical focus. This includes additional transparency dimensions being defined by other documents.

2 Considered Documents

We compared transparency definitions and requirements in several pivotal documents from the field of AI regulation and standardization shown in Tab. 1. Besides the AI Act itself [7], considered sources include central papers with respect to the development of the AI Act, i.e., the HLEG GL [6], and OECD [18], already available documents from standardization, i.e., ISO 22989 [12] and IEEE 7000 or [11], as well as further guidelines and internationally recognized white papers in this direction, i.e., VDE 90012 [22] and Fraunhofer GL [20]. The selection of the documents was based on discussions that were conducted in the context of the German Standardization Roadmap on Artificial Intelligence [4] and its subsequent activities. According to its character as a pilot study, this paper did not include a comprehensive analysis of potentially relevant documents in this field, but focused on this specific selection. In the following, we present the documents in detail.

2.1 HLEG GL: HLEG Ethics Guidelines for Trustworthy AI

The “Ethics Guidelines for Trustworthy AI” by the Independent High-Level Expert Group on Artificial Intelligence of the European Commission [6] is a report published in 2019 defining a European framework for achieving trustworthy AI. These guidelines revolve around ensuring AI systems are lawful, ethical, and robust, commencing from their development to their deployment and operation. They put forward a set of seven key requirements including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, as well as accountability. A notable part of the guidelines is a detailed assessment list designed to guide practical implementation.

■ **Table 1** Overview of the documents considered in this study, by used abbreviation, official title as appears on the document, and corresponding handle to the entry in the references section.

Abbv.	Official document title	Reference
HLEG GL	Independent High-Level Expert Group on Artificial Intelligence (HLEG) set up by the European Commission: Ethics Guidelines for Trustworthy AI	[6]
AI Act	European Parliament legislative resolution of 13 March 2024 on the 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 (Texts Adopted)	[7]
ISO 22989	ISO/IEC 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	[12]
OECD	OECD Framework for the Classification of AI Systems	[18]
VDE 90012	VCIO based description of systems for AI trustworthiness characterisation VDE SPEC 90012 V1.0 (en)	[22]
Fraunhofer GL	Fraunhofer IAIS: Guideline for Designing Trustworthy Artificial Intelligence – AI Assessment Catalog	[20]
IEEE 7000	IEEE Std 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design	[11]

2.2 AI Act: European Artificial Intelligence Act

The “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” [7], commonly referred to as the “European AI Act”, is a comprehensive legal framework on AI devised by the European Union. Representing the first such framework worldwide, it aims to foster trustworthy AI within Europe and beyond by ensuring that AI systems uphold fundamental rights, safety, and ethical standards. Addressing the diverse impact of AI systems, the Act categorizes AI technologies into different risk levels. AI systems carrying an unacceptable level of risk are prohibited. High-risk AI systems face stringent requirements to manage their risks, including issues related to transparency. For limited-risk applications, the AI Act prescribes specific transparency obligations, ensuring an informed and aware interaction with the AI system. Furthermore, the Act outlines particular transparency obligations for all general-purpose AI (GPAI) models and more specific requirements for GPAI models with systemic risk. Additionally, the AI Act imposes stringent obligations on all actors in the AI value chain, ranging from providers and deployers to importers and distributors, among others, ensuring a rigorous approach to enforcement and compliance across the European market.

2.3 ISO 22989

ISO 22989 [12] “Information technology — Artificial intelligence — Artificial intelligence concepts and terminology”, released in 2022, aims to establish a common terminology and concepts for the field of AI with a very general audience.

Terms ranging from “AI agent” to “validation data” are structured into seven categories and defined briefly, usually in single-line descriptions, while the descriptions of concepts span one to several paragraphs each, split into 19 categories. Additionally, elements such as the AI life cycle or AI ecosystems are defined and explained.

2.4 OECD: Framework for the Classification of AI Systems

Based on the first version of the OECD AI Principles [17], the OECD Framework for the Classification of AI Systems [18] is a tool designed to help policymakers, regulators, and legislators characterize AI systems for aligned policy action. This framework examines the spread of AI across sectors, recognizing the variations in benefits, risks, and policy challenges offered by different AI system types. By highlighting system characteristics critical for technical and procedural measure implementation, it aims at facilitating policy debate, supporting risk assessment, and helping in developing AI-related policies and regulations. The framework is structured along five key dimensions, including People & Planet, Economic Context, Data & Input, AI Model, and Task & Output, each with sub-dimensions important for policy considerations. It also distinguishes between AI “in the lab” and AI “in the field,” offering a baseline for promoting common AI understanding, informing AI registries, and supporting sector-specific frameworks, risk assessment, and management throughout the AI system life cycle.

2.5 VDE 90012: VCIO Based Description of Systems for AI Trustworthiness Characterisation, VDE SPEC 90012

The VDE SPEC 90012 “VCIO based description of systems for AI trustworthiness characterisation” by the German Association for Electrical, Electronic & Information Technologies (VDE) [22] provides a framework for describing socio-technical attributes of systems with integrated AI, particularly where high levels of trust are required. It explains the VCIO (Values Criteria Indicators Observables) model, which evaluates a product’s adherence to specific values and its trustworthiness, potentially supporting a trust label certification. This characterization is versatile, serving end consumers, companies, and government entities for setting requirements or comparing products. The assessment allows for different values such as privacy and transparency, and supports tailoring target requirements during product development for value compliance. Notably, while independent of the product’s risk level and without setting minimum standards, the description aligns with the European AI Act, offering a delineation of trustworthiness that demonstrates compliance and market differentiation. Focusing on AI-specific features like datasets, scope, processes, and responsibilities, the standard also encompasses broader elements essential for establishing AI trustworthiness. This VDE SPEC aims to enable a reproducible and transparent classification of AI systems according to their degree of fulfillment of values or competencies, and to allow an assessment of the extent to which the requirements for achieving a certain risk level are met.

2.6 Fraunhofer GL: Fraunhofer IAIS Guideline for Designing Trustworthy Artificial Intelligence

The “Guideline for Designing Trustworthy Artificial Intelligence” released by Fraunhofer IAIS [20] provides a structured approach to define application-specific assessment criteria emphasizing quality and trust as competitive advantages. This guideline is targeted at data scientists in the development stage, and assessors in quality assurance for AI applications. It outlines a four-step assessment process encompassing comprehensive risk analysis, setting measurable targets, listing measures to achieve those targets, and establishing a safeguarding argumentation. The guideline focuses on six dimensions of trustworthiness: fairness, autonomy and control, transparency, reliability, safety and security, and data protection. It includes established KPIs to quantify targets and offers guidance on the documentation of technical and organizational measures reflective of the current state of the art to mitigate AI-related risks.

2.7 IEEE 7000: IEEE Standard Model Process for Addressing Ethical Concerns during System Design

The IEEE Std 7000™-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design [11] was released in 2021 and aims to standardize approaches to consider ethical aspects in the development of systems. As indicated by its title, IEEE 7000-2021 focuses on the development process of the system rather than on properties of the system itself. It is also not exclusively an AI-related standard, describing itself as “applicable to all kinds of products and services, including artificial intelligence (AI) systems”. The five year development process of the standard means that it is not to be understood as a response to very recent advances in generative AI and large language models. The primary audience is “engineers and technologists” viewed from the organization level for whom a set of processes is proposed (including an Ethical Values Elicitation and Prioritization Process, an Ethical Risk-Based Design Process and a Transparency Management Process) to enable them to include ethical aspects into the system development.

In the collection of documents cited here, IEEE 7000-2021 stands out as the only document not primarily conceived for AI applications but for systems development in general, which also reflects onto its perspective of transparency.

3 Comparison Criteria: “Does the document distinctly...”

In this section, we define and delimit our evaluation criteria. The goal of the process was to achieve a comparison that indicates whether or not the documents incorporate or express a specific notion or scope of transparency. Based on the criteria defined below, the particular documents were evaluated. The evaluation had three possible results:

1. yes – the document distinctly adopts the notion of transparency expressed by the criterion;
2. no – the document does not refer to the notion of transparency expressed by the criterion;
3. unclear – the document does not explicitly introduce the notion of transparency expressed by the criterion, but may contain implicit support or alignment with it.

The criteria were developed according to discussions conducted as a follow-up of the German Standardization Roadmap on Artificial Intelligence [4]. This was performed in an iterative approach where relevant aspects were collected from the included documents and the criteria were consolidated accordingly before the evaluation process was started. The authors consider the list of criteria deduced by the analysis of the above-mentioned documents to be exhaustive, meaning that no other additional generic transparency criteria were found in the reviewed documents. The authors do not exclude that new transparency criteria can be deduced, by analyzing further documents. This is part of the iterative approach and the authors recommend screening for additional transparency criteria when expanding this evaluation methodology to an extended set of documents.

For each paper, at least two members of the authors group performed the analysis. In case of a disagreement, the particular rating was discussed in the overall group of authors, who finally decided about the classification. For practical reasons, not all authors could be involved every time. For achieving a reliable consensus, at least five of the authors had to be involved into the final decision where the two authors which analyzed the document needed to be included. In cases where no full agreement could be achieved, the item was assigned to the “unclear” category. Note, however, that the evaluation category “unclear” refers to the way a specific transparency criterion is presented within a document and not to the modalities of agreement among authors.

10:6 Evaluating Dimensions of AI Transparency

The following list describes the used criteria in a systematic way. All criteria start with the phrase “Does the document distinctly...”, followed by the criterion, such as “... set transp. requirements relating to the design or development stage of AI?”.

The term “distinctly” as opposed to “explicitly” is chosen deliberately to include clearly implied intentions, since the choice of words and voice differs considerably between documents. For example, IEEE 7000 [11] states:

System requirements for machine learning systems may include quantitative and qualitative data-oriented specifications that include identifications for collection of data, data formats, diversity, ranges of data, ...

This clearly implies that IEEE 7000 considers transparency requirements relating to the design and development stage of AI, since most commonly for machine learning systems, data will be used for training and testing during these stages – even if no explicit reference to the design and development stage of the life cycle is made. Such clearly implied intentions are, for this study, considered equivalent to explicit statements, with particular decision principles given in the following subsections.

In this process, it should be understood that the task of comparing documents that widely differ in focus, authors’ background, scope and intended impact cannot be strictly formal and performed under sharply defined criteria while still extracting meaningful results. The goal of this approach is to adequately reflect the conceptual ideas incorporated into the respective documents, requiring to some extent a margin of discretion and interpretation. The same should be applied by the readers who are to realize that the following will not replace a formal, thorough study of any individual document, possibly with technical and legal expertise, in particular when, for example, assessing a system with respect to a given standard or legal regulation, such as the AI Act. Furthermore, this motivates the aforementioned choice of three instead of just two possible evaluation results, namely “yes” and “no” when no considerable ambiguity was found, and “unclear” else (cf. Tab. 3).

For the description of the requirements, we use the coding scheme laid out in Tab. 2. For example, the code **TR-STG-OPS** (“Does the document distinctly set transp. requirements relating to the operation stage of AI?”) is composed of the supergroup **TR** for “transparency”, the group **STG** for life cycle stage-related criteria, and the criterion **OPS** for the operation stage.

3.1 ... define the term “transparency” or a closely related concept?

This criterion is satisfied if the document clearly attempts to define or delimit what the term “transparency” means (at least for the specific purpose of the document). Since terms such as “explainability” or “traceability” are frequently used interchangeably, a definition of one of these terms also satisfies the criterion if the document uses the alternative term in a closely related sense.

3.2 ... set transp. requirements relating to the design or development stage of AI?^{TR-STG-DDV}

This criterion is satisfied if the document advocates transparency requirements that must be met or at least considered during the design or development stages of the AI system life cycle, such as the disclosure of training data or AI model details.

■ **Table 2** Abbreviations of the form **TR-XXX-YYY** used here to classify grouped transparency criteria.

Abbrv.	Does the document distinctly...
TR-	<i>Group: Transparency-related (always given here)</i>
STG-	<i>Group: Related to stages within the AI life cycle</i>
DDV	...set transp. requirements relating to the design or development stage of AI?
OPS	...set transp. requirements relating to the operation stage of AI?
EOL	...set transp. requirements relating to post-end of life/retirement/disposal stage of AI?
SYS-	<i>Group: Related to the AI system</i>
MDL	...relate transp. to technical AI properties, such as code or ML models?
WGT	...relate transp. to ML “weights” or “features”?
OUT	...relate transp. to explainability of particular outputs?
CSQ	...relate transp. to predictability of consequences?
LIM	...relate transp. to limits or error/failure modes of AI?
SOA	...limit requirements to a current “state of the art”?
DAT	...relate transparency to training data?
PII	...relate transp. to user data and/or privacy?
PRP	...relate transp. to an intended purpose of the AI?
BIZ	...relate transp. to business models/operator interests?
RVL	...require for transp. revealing to users that the system uses AI as such?
AUD-	<i>Group: Related to the target audience of the transp.</i>
SPC	...consider transparency to be target audience-specific?
DEF	...define one or more such target audiences?
USR	...name users as a target audience?
NNU	...name affected non-users as a target audience?
OPR	...name operators as a target audience?
TST	...name testing and auditing organizations as a target audience?
REG	...name regulators or authorities as a target audience?
DEV	...name developers and (direct) partners in the dev. process as a target audience?
DSA	...name other manufacturers/providers of downstream applications as a target audience?

3.3 ... set transp. requirements relating to the operation stage of AI?^{TR-STG-OPS}

This criterion is satisfied if the document advocates transparency requirements that must be met or at least considered during the operation stage of the AI system life cycle, such as informing users that the system they are using is based on AI.

3.4 ... set transp. requirements relating to post-end of life/retirement/disposal stage of AI?^{TR-STG-EOL}

This criterion is satisfied if the document describes transparency risks or requirements that address the end of life or post-end of life stage of AI, for example measures to be taken during the decommissioning of an AI system, such as assuring and documenting that all data and learned features have been deleted. This criterion explicitly does not refer to transparency problems during operation causing a decommissioning (this would be requirements for the operation stage), but only to transparency-related measures that must be taken once the decommissioning is decided.

10:8 Evaluating Dimensions of AI Transparency

It should be noted that no agreement exists whether the (post) end of life stage is part of the “AI life cycle” in general. Prominently, the OECD [19] explicitly does not include such a stage in its life cycle, stating:

AI system lifecycle: AI system lifecycle phases involve: i) ‘design, data and models’; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) ‘verification and validation’; iii) ‘deployment’; and iv) ‘operation and monitoring’.

A similar subdivision is given in [17] and adopted in [18]. In contrast, ISO 22989 [12] and IEEE 7000 [11] do include this stage, the latter taking its life cycle definition from ISO/IEC/IEEE 12207 [15], as:

2.31

life cycle

evolution of a system, product, service, project, or other human-made entity from conception through retirement

3.5 ... relate transp. to technical AI properties, such as code or ML models?^{TR-SYS-MDL}

This criterion is satisfied if the document states that transparency may include insight into technical system properties such as code, algorithms or ML models – for example by requiring that corresponding design choices must be documented or that the code must be disclosed to certain parties. This criterion does not relate to the machine learning variables in the model – this is covered in **TR-SYS-WGT**. For some parameters (namely “hyperparameters” in machine learning), the distinction is not strict – however, this ambiguity did not arise in the review of the particular documents presented here.

3.6 ... relate transp. to ML “weights” or “features”?^{TR-SYS-WGT}

This criterion is satisfied if the document advocates to disclose, for machine learning-based systems, parameters that were established through model training. These are typically referred to as weights or features.

3.7 ... relate transp. to explainability of particular outputs?^{TR-SYS-OUT}

This criterion is satisfied if the document relates transparency to the provision of explanations for one particular output of the system. For example, in a system that identifies cancer cells in medical images, this could mean to additionally visualize relevant image regions and/or provision of similar reference images to either the professional operator, or the patient.

3.8 ... relate transp. to predictability of consequences?^{TR-SYS-CSQ}

This criterion is satisfied if the document relates transparency to the possibility of predicting and limiting consequences of AI system outputs in its real-world application. This includes transparency with respect to residual risks and potential harms the AI system has. For example, a document may require a company selling an autonomous shuttle to indicate accident risks that can arise from an AI error. We distinguish between this criterion addressing the practical consequences (including long-term and indirect effects) and the criterion **TR-SYS-LIM**, which relates only to a description of immediate technical failure modes and limitations, e.g., in performance, without an actual estimation of the impact associated with the failure.

3.9 ... relate transp. to limits or error/failure modes of AI?^{TR-SYS-LIM}

This criterion is satisfied if the document relates transparency to the disclosure of known limitations and error/failure modes of an AI system. This can include a description of known conditions under which the AI system cannot achieve the required performance (e.g., by specifying the operational design domain, ODD, cf. [2]) or a description of error rates. We distinguish between this criterion addressing the immediate technical failure modes on the technical level, and the criterion **TR-SYS-CSQ**, which relates to a description of practical and possibly physical and/or long-term consequences (primarily on the level of the real-world application).

3.10 ... relate transparency to training data?^{TR-SYS-DAT}

This criterion is satisfied if the document relates transparency of a machine learning system to the disclosure of training, validation or testing data used in the design or development stages of the system – regardless of whether the document advocates disclosing training datasets completely, or documenting selected properties such as fairness or scale of the dataset(s).

3.11 ... relate transp. to user data and/or privacy?^{TR-SYS-PII}

This criterion is satisfied if the document relates transparency to the disclosure of how the AI system utilizes, stores and/or shares user data, privacy-critical information, or personally identifiable information (PII).

3.12 ... limit requirements to a current “state of the art”?^{TR-SYS-SOA}

This criterion is satisfied if the document limits the imposed transparency requirements to what is feasible or accepted based on the current state of the art or the best available techniques (BAT). This implies that the state of the art may evolve and lead to different requirements, but also that acceptance criteria must not be applied in hindsight of later developments. Additionally it implies that the current state of the art is likely to provide an acceptable result at the given time.

3.13 ... relate transp. to an intended purpose of the AI?^{TR-SYS-PRP}

This criterion is satisfied if the document relates transparency to the disclosure of an “intended purpose” of the AI system, if such a purpose exists. The criterion serves to fix the application context as a basis for further development steps like risk management, but also transparency requirements concerning, e.g., which use context needs to be considered when providing information to users. The criterion **TR-SYS-PRP** may serve to convey system capabilities to users and avoid misunderstandings about its proper use. Beyond this, the disclosure may serve to reveal hidden interests – however, the particular case of hidden *business interests* is addressed by **TR-SYS-BIZ** specifically.

3.14 ... relate transp. to business models/operator interests?^{TR-SYS-BIZ}

This criterion is satisfied if the document relates transparency to the disclosure of business interests of the suppliers or operators of the AI system, or similar interests for providing or operating the system. For example, the document may advocate that a company offering an app for health advice must disclose to users if their business interest is to build a general human health prediction model for insurance companies – even if the individual users’ privacy

10:10 Evaluating Dimensions of AI Transparency

is believably protected in the process. Note that this criterion goes beyond **TR-SYS-PRP** in the sense that a document satisfying **TR-SYS-BIZ** must distinctly consider business interests (e.g., the collection of health data) to extend beyond the immediate purpose (e.g., the provision of a health advice app). To satisfy this criterion thus also means to acknowledge a possible conflict between transparency and business interests.

3.15 ... require for transp. revealing to users that the system uses AI as such?^{TR-SYS-RVL}

This criterion is satisfied when the document requires or proposes for the benefit of transparency that users should be informed about the fact that the system uses or is based on AI, or that its output is (at least partially) AI generated. This criterion is only satisfied if the document clearly considers the fact itself important to reveal proactively. It is not, for example, satisfied if the document requires an AI system to adhere to regulations by which an interested user may come to know that it uses AI, for example by registering in an official database.

3.16 ... consider transparency to be target audience-specific?^{TR-AUD-SPC}

This criterion is satisfied if the document indicates that transparency cannot always be defined universally, but instead should be evaluated with respect to, or designed for, a particular target audience. For example, this audience may be defined through professional experience, (lack of) AI literacy, personal handicaps, the specific role in the AI value chain respectively the usage of the AI system, or the level information relevant to the particular audience. This criterion is only satisfied if the document indicates that different levels of information or different means of presentation are adequate for different audiences. Merely listing different receiving groups or entities will not satisfy this criterion. Furthermore, the document must clearly suggest providing multiple such levels of information regarding the same system. A document that acknowledges the existence of different levels of expertise, but derives from this the requirement to release a single documentation that must be understandable to all stakeholders alike (e.g., by assuring all explanations can be understood by laypersons) will not satisfy this criterion.

3.17 ... define one or more such target audiences?^{TR-AUD-DEF}

This criterion is satisfied if the document considers transparency to be target audience-specific (**TR-AUD-SPC**) and, in addition to this, names or defines concrete groups that may require different variants of transparency. It must not define detailed requirements for transparency, but it should specifically mention one or more target audiences with particular requirements.

3.18 ... name users as a target audience?^{TR-AUD-USR}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, the end users of the AI system.

3.19 ... name affected non-users as a target audience?^{TR-AUD-NNU}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, people affected by the AI system who are not actively involved in their operation (such as operators or users) in the sense of the economic concept of “negative externalities” [9, Chapter 1, p. 5; Chapter 5, p. 125 ff.].

Prototypical examples include pedestrians, who are affected by the operation of an AI-based automated road vehicle; or job candidates whose application documents are rated through an AI-based system used by recruiters.

3.20 ... name operators as a target audience?^{TR-AUD-OPR}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, the operators of the AI system when they differ from the users. We use the term “operator” in the sense of the “natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity”.¹ In particular, the “operator” in this case is the body directly responsible for the continued operation of an AI system.

3.21 ... name testing and auditing organizations as a target audience?^{TR-AUD-TST}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, officially appointed public or private testing and auditing organizations, registration centers, etc., who are responsible for performing particular tests of the AI system, certifying it, providing an operating license, or similar. This includes notified bodies or other conformity assessment bodies included into the certification or conformity assessment of an AI system.

3.22 ... name regulators or authorities as a target audience?^{TR-AUD-REG}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, regulators, authorities, legislators, or other public bodies responsible for controlling the development and operation of AI systems.

3.23 ... name developers and (direct) partners in the dev. process as a target audience?^{TR-AUD-DEV}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, persons or organizations who are involved in the development or production of the AI system. This includes developers and other departments in the own company, but also partners or suppliers along a supply chain who are involved in the development process and related areas.

3.24 ... name other manufacturers/providers of downstream applications as a target audience?^{TR-AUD-DSA}

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, other manufacturers or providers who intend to use the system in a downstream application. That is, these entities modify, extend, or adapt the original system and thus, create a new system, e.g., with a modified scope

¹ This definition is verbatim from the AI Act [7], which, however, assigns the term “deployer” to it; whereas the AI Act regards “operator” as “a provider, product manufacturer, deployer, authorised representative, importer or distributor.”

10:12 Evaluating Dimensions of AI Transparency

■ **Table 3** Results of the comparison. ✓ indicates that the document distinctly adopts the concept of the criterion; — indicates that it does not. Cases where the distinction is unclear are marked with ○.

	HLEG GL	AI Act	ISO 22989	OECD	VDE 90012	Fraun- hofer GL	IEEE 7000
Definition	✓	✓	✓	○	○	○	✓
TR-STG-DDV	✓	✓	✓	✓	✓	✓	✓
TR-STG-OPS	✓	✓	✓	✓	✓	✓	✓
TR-STG-EOL	—	—	✓	○	○	—	—
TR-SYS-MDL	✓	✓	✓	✓	✓	✓	○
TR-SYS-WGT	—	—	✓	✓	○	✓	—
TR-SYS-OUT	✓	✓	✓	✓	✓	✓	—
TR-SYS-CSQ	✓	✓	—	✓	○	✓	✓
TR-SYS-LIM	✓	✓	✓	○	○	✓	○
TR-SYS-SOA	✓	✓	—	—	○	✓	✓
TR-SYS-DAT	✓	✓	✓	✓	✓	✓	✓
TR-SYS-PII	○	✓	✓	✓	✓	✓	✓
TR-SYS-PRP	✓	✓	✓	✓	—	✓	✓
TR-SYS-BIZ	✓	—	○	○	—	✓	✓
TR-SYS-RVL	✓	✓	—	✓	✓	✓	○
TR-AUD-SPC	✓	✓	✓	✓	✓	✓	✓
TR-AUD-DEF	✓	✓	—	✓	○	✓	—
TR-AUD-USR	✓	✓	✓	✓	✓	✓	✓
TR-AUD-NNU	✓	✓	—	✓	✓	✓	✓
TR-AUD-OPR	○	✓	○	✓	✓	✓	✓
TR-AUD-TST	✓	✓	○	—	✓	✓	○
TR-AUD-REG	—	✓	○	—	✓	—	—
TR-AUD-DEV	○	✓	✓	—	✓	✓	✓
TR-AUD-DSA	○	✓	—	—	—	○	—

or intended purpose. For example, this applies when the manufacturer of a downstream application builds its system on a general purpose AI (GPAI) model by adapting this base model towards a specific use case. The question here is whether the provider of the GPAI model is requested to provide certain transparency information to the manufacturer of the downstream application. This criterion, based on the concept of GPAI, considers “downstream applications” to be potentially unforeseen by the developers of the original AI system; in contrast, **TR-AUD-DEV** refers to informing parties working along the same known supply chain.

4 Results

Table 3 provides an overview of our analysis of criteria for AI transparency dimensions across the documents described above. The key findings from our review are presented, subsequently. For each entry, a final consensus could be found in the group of authors, according to the defined evaluation process. In some cases, the discussions showed that an ambiguity in the

interpretation of the requirements persisted. Thus, the particular criterion had to be rated as “unclear”, since the definitions in the corresponding document were not clear enough or could not be interpreted in a sufficiently consistent way.

Transparency is a subject of discussion in all of the documents, emphasizing its importance in AI development. However, in some documents its definition is unclear or not provided, i.e., in OECD, VDE 90012, and the Fraunhofer GL.

Explainability of system outputs is included as an essential dimension of transparency in nearly all the documents. The only document not mentioning this aspect, namely IEEE 7000, is also the sole source in our collection not specific to AI but to system design in general. In certain documents, such as the Fraunhofer GL, HLEG GL, and the AI Act, traceability and communication are additionally mentioned as integral parts of transparency.

Transparency is explicitly demanded during both the development and operational stages across nearly all documents, except for OECD. The end of life stage, in particular in terms of transparency obligations as a task in the retirement stage, are only made explicit in ISO 22989. Furthermore, no concrete measures are expressed for this stage, even when this stage was included in the document. As previously stated, it should also be noted that the inclusion of this stage as part of the life cycle is disputed.

All documents claim that transparency requirements must be target group-specific. However, the definition of target groups is often unclear or limited.

There are significant differences regarding which target groups and use cases are considered. End users are consistently included, and also indirectly affected stakeholders / non-users are explicitly mentioned in nearly all documents, with the exception of ISO 22989. Transparency towards operators is widely addressed as well, although not always explicitly, i.e., in the HLEG GL and ISO 22989. Transparency towards developers is absent in OECD and addressed unclearly in the HLEG GL. For the other documents, adequate information needs to be provided to the developers in order to achieve sufficient transparency. Transparency for testing or auditing organizations is explicitly mentioned just in four out of seven documents, while transparency for regulators and public authorities is only addressed in a dedicated way in the AI Act and VDE 90012. More complex scenarios for the implementation of AI systems involving different actors are usually not considered. This means that most of the documents refer to a situation with a single manufacturer of the AI system. Only the AI Act explicitly discusses more complex scenarios, where, e.g., a general purpose AI (GPAI) model is developed by one manufacturer and then integrated into a downstream application by another manufacturer. In these cases, the AI Act includes transparency obligations to be fulfilled by the manufacturer of the GPAI model.

Notable differences were recognized regarding the specific kind of information that should be made transparent. Transparency concerning code, models, and training data is consistently addressed across documents. As already mentioned, transparency with respect to the explanation of outcomes is required in most of the documents, except IEEE 7000. Similarly, information regarding an AI system’s intended purpose and its predictable consequences is considered a significant transparency aspect in nearly all the documents, with the exception of VDE 90012 regarding the intended purpose and of ISO 22989 regarding the predictability of consequences. Information concerning the limitations and error modes of AI systems is generally addressed in the examined sources, even though some documents do not explicitly present this as a dimension of transparency. Finally, other aspects such as the reference to the state of the art, to the business model and operator interests, as well as the disclosure of technical parameters of the system (e.g., “weights” and “features”) are not mentioned as transparency obligations in many of the documents.

Most of the documents also require that an AI system should reveal that the user is interacting with an AI system. This requirement is only absent in ISO 22989 and unclear in IEEE 7000. Finally, it can be recognized that many entries remain marked with a grey circle. This shows that a number of topics remains ambiguous in the documents.

5 Discussion

The presented results underline that transparency is regarded as an important ethical value in AI regulation and standardization literature. Also, there is a wide range of criteria that are considered as important requirements for AI transparency across different regulatory frameworks and guidelines. This variety of criteria reflects the fact that AI transparency is thought to have many dimensions, each of which demanding specific, tailored requirements to be addressed. Even though many transparency dimensions address overlapping technical or societal aspects, they allow for independent implementation of transparency measures. For example, informing users that they are engaging with an AI system, explaining to auditing organizations or authorities how the system processes inputs to reach outputs, and providing access to information about the AI model and training data can be executed separately using distinct approaches.

When considering the full extent of what defines transparency within the respective documents, considerable disparities become evident. No two documents share a completely compatible conception of this ethical value. A contributing factor is that the considered documents define the criteria for transparency – to a large degree – implicitly through the measures proposed, rather than comprehensively specifying concepts and goals first and deriving adequate measures systematically. Moreover, different kinds of documents have different scopes and focuses; therefore they tend to emphasize transparency aspects that are more relevant for their purposes. For instance, some of them include more technical specifications regarding appropriate measures of system transparency, while others, like the EU AI Act, focus more on high-level requirements related to AI transparency, explicitly leaving the task of defining concrete technical measures to domain-specific, technical standards.

The analyzed documents reflect that transparency basically deals with the question concerning which pieces of information should be provided to which stakeholder to deploy AI safely and responsibly. Indeed, transparency is generally considered a driver for trustworthiness, enabling meaningful and informed human interaction with the system. In order to do so, it is necessary to consider the system's intended purpose, which defines the specific technical solutions, user groups, and use cases for the AI system. The provision of information needs to be aligned with the expertise of the involved persons as well as their particular needs. For example, there is a considerable difference between laypersons, technical experts, and actors with specific domain, application or regulatory knowledge. Additionally, the adequacy of information depends on the particular role of the actors. For example, a developer needs different information in comparison to a user, provider, auditing organization, or public authority.

This also goes for cases where a manufacturer includes an AI model from a third party provider in a new downstream application. In particular, such more complex scenarios were intensively discussed in the legislative phase of the AI Act [7] in the context of AI models without a specific / narrow intended purpose. The AI Act defines these types of models as general purpose AI (GPAI) models and describes obligations, which information has to be made available by the third-party provider of the GPAI model in order to achieve sufficient transparency for the manufacturer of the downstream application. This substantially extends

the requirements, since the ethical impact can usually only be rated when the context of the application is clear. Thus, a new line of discussions was addressed in the AI Act in this regard compared to the other documents. This was due to the fact that most of these documents were older, i.e., written between 2019 and 2022, where high-impact large language models and other generative AI systems were not yet on the market.

The latest developments of the debate around transparency for GPAI models exemplify the highly dynamic nature of AI technology, which is leading to challenges regarding AI standardization and regulation. As an ethical value guiding the deployment of AI systems, transparency is a fundamental prerequisite for the achievement of other ethical principles. Indeed, a lack of awareness of the system impact and a limited understanding of its capabilities and limitations could hinder different actors along the AI value chain in the responsible use of a system, causing ethical issues ranging from safety risk to data misuse or group discrimination. Moving from these premises, the development of the AI Act [7] and the associated New Legislative Framework of the European Union are indicative of the goal to establish clearly-defined, internationally accepted rules and standards to enable ethical, value-centered use and implementation of AI systems. To this end, the exact definition of major terms, the operationalizability of concepts, and an adequate translation into concrete requirements for all relevant stakeholders along the AI value chain plays a pivotal role. The dimensions and criteria presented in the current paper are considered to serve as a starting point for systematically collecting and comparing requirements for AI transparency. In particular, this may apply to the identification of discrepancies in current standardization documents and guidelines as well as to a systematic compilation of key aspects in future development steps.

At the same time, the presented study has limitations to be considered. For practical reasons, the study was performed as a pilot study and limited to seven well-known documents in the AI context. However, a broader look that also includes scientific positions could help to establish a more thorough perspective in this very dynamic field. Additionally, the currently developed standards that include aspects of transparency should be addressed using the presented approach. For example, this may refer to important standards like ISO/IEC 42001 [13] regarding management systems for AI, or ISO/IEC 12792 [14], which directly addresses a taxonomy for transparency of AI systems. Furthermore, the development towards harmonized standards and guidelines for the implementation of the AI Act should be taken into account.

It must be noted that the considered and compared documents were released in the time span between 2019 [6] to 2024 [7]. These past five years have been characterized by an unprecedented level of disruption in AI technology, such as the development of large language models and generative AI models for image and video synthesis. These new types of models blur the line between human and AI capabilities – accompanied by a substantial shift in the perception of AI systems, their potentials, societal impacts and risks, and regulatory requirements. Hence, the documents, even though seemingly released in quick succession, must already in part be viewed in their individual “historical” context, explaining, for example that the AI Act [7] heavily addresses the challenges of GPAI, while in IEEE 7000 [11], many transparency requirements common for machine learning are absent. This explains in part the heterogeneity of the analyzed documents, such as concerning requirements with respect to GPAI models.

Further on, our analysis encountered challenges due to the inherent complexity and occasional ambiguity of the sources. The fulfillment of specific criteria was not always explicitly stated in a single, easily identifiable section, necessitating an interpretative evaluation of the

overall narrative of the document. To maintain objectivity and ensure the accuracy of our classifications, our research team implemented a systematic review process. This involved meticulous discussion and careful cross-referencing of the documents to ensure our evaluations were anchored in the text. This systematic approach, guided by the framework set out in our methodology, aimed to limit the influence of subjective judgments and provided an accurate representation of each document's stance on transparency requirements, as depicted in Tab. 3.

6 Conclusion and Outlook

In conclusion, our study offers an initial framework for evaluating and comparing ethical concepts within standards and regulations, specifically focusing on the notion of transparency in AI. Its application has highlighted that the definitions of transparency differ to a considerable degree among the considered documents. These differences can, to some extent, be attributed to the different purposes of these documents.

However, in the absence of a standardized framework for representing and comparing these definitions and their differences, the variety of notions arguably hinders the underlying goal of establishing a shared understanding across stakeholders.

Future research should expand upon this approach by assessing the criteria for transparency against a broader spectrum of references, which should include documents in the field of standardization on the one hand, e.g., regulations, standards, or guidelines; and on the other hand key academic publications as well as high-impact white papers.

In the European context, the implementation of the AI Act represents a particularly promising case for the application of our methodology as it necessitates the formulation of more specific and detailed standards for AI. By assessing how the various concepts and requirements for transparency adopted in different documents align with those found in the AI Act, our approach can help to identify key areas for creating a harmonized regulatory framework. These efforts should not only be consistent with the AI Act but also seek to elaborate more detailed standards for sector-specific applications.

Ultimately, our approach is designed to create clear and precise definitions for transparency dimensions and corresponding operationalizable criteria, representing the building blocks for agile AI governance frameworks. By doing so, we equip policymakers and industry stakeholders with fundamental tools to ensure AI trustworthiness and enhance their capacity to adapt to the ongoing developments in technology and the evolution of ethical standards in the AI field.

References

- 1 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020. doi:10.1016/J.INFFUS.2019.12.012.
- 2 ASAM e.V., Advanced Data Controls Corp., Ansys Inc, AVL List GmbH, BTC Embedded Systems AG, DENSO Corporation, Deutsches Zentrum für Luft- und Raumfahrt e.V., Edge Case Research, e-SYNC Co. Ltd., FIVE, Foretellix Ltd, Fraunhofer-Institut für Kognitive Systeme IKS, Hexagon Manufacturing Intelligence, iASYS Technology Solutions Pvt. Ltd, Institute of Communication and Computer Systems (ICCS), Oxfordshire County Council, RISE Research Institutes of Sweden, Robert Bosch GmbH, Siemens Digital Industries Software, SOLIZE Corporation, Technische Universität Braunschweig Institut für Regelungstechnik, and WMG University of Warwick. ASAM OpenODD: Concept Paper, October 2021.

- 3 China Academy of Information and Communications Technology (CAICT). White Paper on Trustworthy Artificial Intelligence, 2021.
- 4 DIN, DKE. German Standardization Roadmap on Artificial Intelligence (2nd edition), 2022.
- 5 Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.
- 6 European Commission / Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019.
- 7 European Parliament and the Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act), June 2024. Official Journal of the European Union L 218, 12.7.2024. ELI: <http://data.europa.eu/eli/reg/2024/1689/oj>.
- 8 Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1):2053951719860542, 2019. doi:10.1177/2053951719860542.
- 9 J. Gruber. *Public Finance and Public Policy, Fifth Edition*. Worth Publishers / Macmillan Learning, 2016.
- 10 Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020.
- 11 Institute of Electrical and Electronics Engineers (IEEE). IEEE 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design, 2021.
- 12 ISO/IEC. ISO/IEC 22989: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, April 2021.
- 13 ISO/IEC. ISO/IEC 42001: Information technology — Artificial intelligence — Management system, December 2023.
- 14 ISO/IEC. ISO/IEC DIS 12792: Information technology — Artificial intelligence — Transparency taxonomy of AI systems, July 2024.
- 15 ISO/IEC/IEEE. ISO/IEC/IEEE 12207:2017: Systems and software engineering — Software life cycle processes, November 2017.
- 16 National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
- 17 Organisation for Economic Co-operation and Development (OECD). Scoping the oecd ai principles, 2019. doi:10.1787/d62f618a-en.
- 18 Organisation for Economic Co-operation and Development (OECD). OECD Framework for the Classification of AI Systems, 2022.
- 19 Organisation for Economic Co-operation and Development (OECD). OECD/LEGAL/0449: Recommendation of the Council on Artificial Intelligence, 2023.
- 20 Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B. Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss, and Stefan Wrobel. *Guideline for Designing Trustworthy Artificial Intelligence: AI Assessment Catalog*, February 2023.
- 21 Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22, 2019. doi:10.1007/978-3-030-28954-6_1.
- 22 VDE Verband der Elektrotechnik. VDE SPEC 90012: VCIO based description of systems for AI trustworthiness characterisation, April 2022.

Transparency of AI Systems

Oliver Müller^{1 2} ✉

Federal Office for Information Security (BSI), Saarbrücken, Germany

Veronika Lazar¹

Federal Office for Information Security (BSI), Saarbrücken, Germany

Matthias Heck

Federal Office for Information Security (BSI), Saarbrücken, Germany

Abstract

Artificial Intelligence (AI) has now established itself as a tool for both private and professional use and is omnipresent. The number of available AI systems is constantly increasing and the underlying technologies are evolving rapidly. On an abstract level, most of these systems are operating in a black box manner: only the inputs to and the outputs of the system are visible from outside. Moreover, system outputs often lack explainability, which makes them difficult to verify without expert knowledge. The increasing complexity of AI systems and poor or missing information about the system make an assessment by eye as well as assessing the system's trustworthiness difficult. The goal is to empower stakeholders in assessing the suitability of an AI system according to their needs and aims. The definition of the term transparency in the context of AI systems represents a first step in this direction. Transparency starts with the disclosure and provision of information and is embedded in the broad field of trustworthy AI systems. Within the scope of this paper, the Federal Office for Information Security (BSI) defines transparency of AI systems for different stakeholders. In order to keep pace with the technical progress and to avoid continuous renewals and adaptations of the definition to the current state of technology, this paper presents a technology-neutral and future-proof definition of transparency. Furthermore, the presented definition follows a holistic approach and thus also takes into account information about the ecosystem of an AI system. In this paper, we discuss our approach and proceeding as well as the opportunities and risks of transparent AI systems. The full version of the paper includes the connection to the transparency requirements in the EU AI Act of the European Parliament and council.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases transparency, artificial intelligence, black box, information, stakeholder, AI Act

Digital Object Identifier 10.4230/OASICS.SAIA.2024.11

Category Practitioner Track

Related Version A full version of the paper is available at the BSI website:

German version: **Transparenz von KI-Systemen**

English version: **Transparency of AI systems**

Acknowledgements We would like to thank our colleagues from the Central Office for Information Technology in the Security Sector (ZITiS) as well as from the Federal Office for Information Security (BSI) for their critical comments and for proofreading the full version of the paper.

¹ These authors contributed equally to this work.

² Corresponding author.



© Oliver Müller, Veronika Lazar, and Matthias Heck;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 11; pp. 11:1–11:7

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

In this paper, we present a definition of transparency for information technology systems that have integrated artificial intelligence (AI). The aim of this publication is to develop a common understanding of the term transparency and to highlight the relevance of transparency for various stakeholders and the BSI. Therefore, the paper is addressed to all stakeholders of AI systems and is intended to show, among other things, that different stakeholders may also have different transparency requirements.

2 Definition

► **Definition 1.** *Transparency of AI systems is the provision of information about the entire life cycle of an AI system and its ecosystem. Transparency promotes accessibility to information that enable an assessment of the system with regard to different needs and objectives for all stakeholders.*

2.1 Elements

The above definition is based on the presentation of the transparency concept in [5] and [2]. It is compliant with the transparency requirements in the EU AI Act (see full version of the paper for details) and represents the position of the BSI. In the following subsections, the individual elements of the definition are described in more detail.

2.1.1 AI system

The EU AI Act governing the regulation of artificial intelligence, defines an AI system as “a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” (cf. Article 3 EU AI Act). In its definition, the BSI explicitly formulates the hardware component and defines AI systems as software and hardware systems that utilise artificial intelligence in order to behave “rationally” in the physical or digital dimension. Based on their perception and analysis of their environment, these systems act with a certain degree of autonomy in order to achieve certain goals [3]. The technology integrated into these systems, known as AI, consists of different disciplines like machine learning, inference and robotics. These include expert systems and neural networks used in AI systems. This is not an exhaustive list of the techniques used, but is intended to show the abundance of different techniques. An equally wide range becomes clear when considering the different functionalities of AI systems, ranging from simple to highly complex tasks. AI systems can perform tasks such as pattern recognition, classifications, forecasts, recommendations, natural language processing or computer vision and can also be combined with each other in a variety of ways. In addition, AI can be implemented in the systems in different ways depending on the requirements and objectives of the application. On the one hand, it can be developed and used as a separate application and represent the primary function of the system, as is the case, for example, in chatbot applications. On the other hand, it can also be integrated into existing systems, for example in order to expand their functionality and/or increase performance in background processes. Considering the way AI systems are implemented, the degree of automation can also vary greatly. While some systems only use the output of AI as recommendations and

require humans as the final decision-making authority, other (sub-)systems autonomously implement the decisions and classifications of AI without further human action. Overall, it can be summarised that there is not one AI system, but rather a wealth of different techniques, functionalities and forms of implementation.

2.1.2 Ecosystem

In this paper, the term ecosystem refers to the context in which an AI system is developed, deployed and operated. The information regarding the ecosystem of an AI system goes beyond the actual AI system and should, for example, also include details about the provider (e.g. location, contact details) or the development process of the system. The term should also include the entire supply chain of the AI system. The decision to include information about the ecosystem of an AI system in the definition of transparency is based on the fact that there is a (conditional) dependency between the actual AI system and its ecosystem. For example, if the AI system is developed and operated outside the European Union in a third country, corresponding questions and challenges arise with regard to the underlying level of IT security and data protection level. This meta-information can support a well-founded assessment of the situation as well as an informed decision by stakeholders.

2.1.3 Information

Information is the basis for the knowledge needed by stakeholders to form an assessment of the AI system and its ecosystem. They must be disclosed and made accessible so that they are available for this purpose. In addition, information must be relevant and appropriate for gaining knowledge. The full version of the paper explains the transparency requirements in the EU AI Act and sets out the minimum information to be disclosed by providers and operators of certain AI systems.

2.1.4 Life cycle

According to [1], the life cycle of an AI system comprises various phases:

- Planning and conception phase
- Design, development and validation phase
- Commissioning and application phase
- Continuous evaluation phase
- System updates
- Decommissioning

A brief description of the phases in the context of the concept of transparency is given in the full version of the paper.

2.1.5 Needs and objectives

These are individual and can be quite different in varying applications. This term is intended to reflect the fact that transparency is not intended to enable access to specific information, but rather to provide information that enables the respective stakeholders to make an assessment. What is specifically assessed by the respective stakeholder is individual and contextual. Stakeholder needs and objectives vary depending on the application scenario. The aim is to cover a wide range of desirable system information enabling the system to be assessed in terms of the specific needs of stakeholders.

2.1.6 Stakeholder

The term “stakeholder” refers to all parties who are either indirectly (e.g. through impacts) or directly (e.g. through application) affected by an AI system or who interact with the system (e.g. developers). These can be individual persons or groups of people. A stakeholder does not have to play an “active” role. While consumers and users usually only use an AI system, it is possible that experts, developers and companies/organisations provide an AI system in addition. Indirectly affected persons/third parties do not provide an AI system, nor do they use it. Nevertheless, they can be affected by the impact and thus become (passive) stakeholders. The presented list of different stakeholders does not claim to be exhaustive and can be refined as desired. However, the chosen representation is sufficient to show that there may be different interests with regard to an AI system, which can be reflected, among other things, in different requirements for transparency – such as the type or level of detail of the information provided – of an AI system. Therefore, the various stakeholders must be taken into account when defining the concept of transparency.

3 Discussion

3.1 Approach and procedure

The existence of already published definitions of the concept of transparency raises the question of the need for a further definition. The sheer breadth of the topic of transparency on the definition market ultimately reflects the different requirements for transparency depending on the stakeholders and the area of application. In order to provide a basis for further work of the BSI with a focus on broad stakeholder groups and generic areas of application, it was decided not to use existing definitions. In addition, the speed at which technologies in the AI sector are developing is enormous. This harbours the risk that definitions, once established, may lose their validity, especially if they are too specific. In order to keep pace with technical progress and to avoid constant renewal and adaptation of the definition to the current state of the art, the definition of transparency presented in this paper is as technology-neutral and future-proof as possible. On the one hand, it should be easy to understand, cover all relevant aspects of transparency and at the same time be open enough to allow individual interpretation depending on the respective stakeholder and the AI technology used. On the other hand, it should serve as a generic basis for future work of the BSI in this area. Furthermore, a holistic approach was taken to the definition: transparency includes both the provision of information about the AI system itself and about its ecosystem, such as the supply chain of the AI system or details about the provider. The weighting of the information provided is the responsibility of the respective stakeholder.

3.2 The aim of transparency

By promoting the transparency of AI systems, the aim is to strengthen the autonomy of stakeholders and enable them to decide for themselves whether the use, modification or provision of an AI system is appropriate and justifiable for them. It is not enough to simply describe the capabilities of the AI system. The limitations of the system must also be analysed and made transparent. Only in this way can a holistic assessment be made by the stakeholders, e.g. whether an AI system is suitable for a particular purpose or not. In the area of digital consumer protection, this should help to ensure that consumers can recognise and use safe and trustworthy AI systems despite increasing digitalisation. Companies and organisations should also be enabled to develop and operate their own AI

systems transparently. Transparent information from the ecosystem of an AI system should also enable third parties affected by impacts to recognise how they can assert their rights in the event of damage. The transparency of AI systems thus serves to empower stakeholders.

3.3 Opportunities through transparency

The use of transparent AI systems can promote the traceability of decisions and the assessment of the appropriateness of systems. Transparency can also help to protect against misuse by enabling potential risks and undesirable effects to be recognised at an early stage. In order to react appropriately to problems, it is important to know, for example, whether the output of an AI system is free from discrimination or whether it violates licence conditions. In terms of consumer protection transparency can also act as a support tool. Transparency provides the basis for a correct assessment of the appropriateness of the system used. In order to be able to make such an assessment at all, information about the system must be accessible. A valid assessment of the adequacy of the AI system forms the basis for positive trust and acceptance processes. Initial publications show, for example, higher download numbers for transparent AI models, which could be an indication of better acceptance of these systems among developers [4]. Lack of transparency complicates the valid assessment of the adequacy of a system, and thus the assessment of its trustworthiness. The latter is a prerequisite for establishing and maintaining a positive relationship of trust with the system and the related output. In addition, transparency can enable users to exercise rights more easily by requiring transparency accompanied by clearer definitions of legal responsibilities and identification of those responsible for the use of AI systems. The listed aspects of traceability, abuse protection, acceptance and trustworthiness as well as legal responsibility show the relevance of transparency in the use of AI systems. This relevance is also reflected in regulatory and legal requirements (see full version of the paper). Transparency can, on the one hand, contribute to the security of AI systems and, on the other hand, promote the safe use of AI applications. In this way, transparency can enable the identification of possible problems and vulnerabilities, make undesirable system behaviour visible and contribute to problem identification and the prevention of misuse. In the context of IT security, transparency also provides the basis for the disclosure and assessment of risks associated with the use of the system. The identification of roles and responsibilities in the event of damage and adverse events as part of transparency requirements can also help stakeholders to detect faulty system behaviour, reduce response times and thus mitigate possible consequential damage. If transparency is practised in the early phases of the life cycle of an AI system, inconsistencies can be avoided within the development team from the outset, sources of error minimised and training phases shortened. AI systems are both developed and increasingly used as development tools – e.g. in AI-supported programming for the automatic generation of program code – for new systems. In the early development phases, it is also important from a developer's point of view to know where the training/test/validation data come from, how they are obtained and whether they are free of bias – e.g. to avoid discrimination. This information is important in order to be able to prepare the data correctly before training/testing/validating (pre-processing). Current development trends also show that pre-existing models are often used, which makes the presence and accessibility of all safety-critical information on existing models particularly relevant. In the absence of this information, there is a risk of implementing the security risks of the basic models into your own products. Both systems are then interdependent, and any lack of transparency in the underlying system is transferred to the system built on top of it. The inheritance of the security risks described above from different areas leads to an increased overall risk in the examples mentioned, which underlines once again

the explosiveness of transparency for security-relevant aspects when using AI systems. In addition, as discussed in the previous section, transparency can contribute to better user empowerment in assessing AI systems. A correct assessment of the application, suitable application scenarios and possible problems and security risks can promote safe use by users.

3.4 Dangers of transparency

So far, the positive aspects of transparency have mainly been presented. Increasing/improving the transparency of AI systems can also have unintended negative effects. For example, the provision of information on the functionality or architecture of an AI system can reveal new attack vectors that attackers can exploit to misuse or compromise the system. Information about limitations or excluded areas of application of an AI system could also be deliberately exploited by attackers, e.g. to deliberately generate erroneous behaviour or destructive output. Conversely, attackers can also abuse the trust that transparency is supposed to create in order to deliberately provide incorrect information. For example, in reality, safety-critical applications can be presented as uncritical. In addition, non-transparent systems can also be marked as transparent. Such pseudo-transparency can be used for marketing purposes of the own product and lead to a wrong assessment of the system by consumers if the transparency label is not checked. Therefore, questions about the trustworthiness of the information disclosed/provided must also be answered in the future. An official transparency label and verifiable transparency criteria could provide a remedy here. The transparency of AI systems is therefore a double-edged sword and should hence be used with caution. The goals and problems are sometimes contradictory and cannot be solved simultaneously. Answering the key questions “What information does a stakeholder need to make a decision?” and “What information is not relevant?” can be helpful. Similar to the EU General Data Protection Regulation (GDPR) the principle of data minimisation is also recommended here, which is answered separately for each individual use case: as much information as necessary, but no more than strictly required, should be disclosed. This “need-to-know” principle applies especially to safety-critical information. The goal should be an appropriate level of transparency, which is sufficient, and at the same time aspects such as security should not be too disadvantaged.

4 Conclusions

Due to the black box properties of many AI systems, data and information are processed in a way that is not transparent to users leading to an unverifiable decision being produced. A lack of knowledge about the AI system goes hand in hand with a lack of traceability and verifiability of system outputs. It is difficult to assess whether system outputs are correct and appropriate. Similarly, questions about responsibility, liability or fairness cannot be answered, if there is a lack of information about the system and its ecosystem. Ultimately, non-transparent AI systems can lead to a loss of trust and a rejection of the system. The implementation of AI components into existing systems and the combination of different systems can also increase complexity and makes it even more difficult to access relevant information. The problems caused by a lack of system insight and a lack of information about the system are manifold and represent a major challenge. Transparency addresses this problem and aims to make AI systems more comprehensible by increasing accessibility to system information and enable a valid assessment of the systems. For these reasons, transparency plays a crucial role for all stakeholders of an AI system. The challenge is to serve all stakeholders with their individual and different transparency requirements. The overall

project and future work on the transparency of AI systems are aimed at all BSI stakeholders. The relevance of the topic for society as a user of the systems is reflected in the expected higher traceability, better protection against abuse, more valid acceptance and trustworthiness processes as well as a more binding legal responsibility. The transparency measures are intended to contribute directly to the empowerment of end users by increasing their trust and autonomy regarding the choice and use of AI systems. Overall, this empowerment of end users aims to democratise the use of AI systems. In addition, the work in the field of transparency and the derivation of concrete criteria and measures should contribute to the overarching goal of the trustworthy use of AI systems. For companies involved in the development of AI systems, the relevance of the topic and the observance of measures in the development and operation of AI systems should be accelerated. Guidelines and positions are to be made available as guidance for stakeholders from the economic environment who want to use third-party AI systems in their organisations or implement them in their systems and products. These guidelines are intended to make it easier for companies to identify suitable, secure and high-performance systems. This work is also intended to provide guidance for public authorities wishing to use AI systems. In addition to their own use, the daily new safety-relevant findings on AI systems, which have to be addressed, pose the challenge of ensuring a technically qualified and adequately positioned staffing level for public sector stakeholders and administrations. This and future work in the field of transparency can be used to facilitate and accelerate permanent and adequate (post-)training of staff. In addition, the establishment of transparency criteria hoped for by this and future work can facilitate the development of meaningful and reliable quality seals by public authorities. With regard to the expected further increasing prevalence and widespread roll-out of AI systems in many areas of life, the relevance of AI systems to society as a whole is steadily increasing. In order to be able to make competent and valid assessments of these systems in the future, the establishment of transparency criteria is indispensable. For providers and operators of certain AI systems – such as general-purpose AI systems or emotion recognition systems – transparency obligations are already defined in the EU AI Act (cf. Article 50 EU AI Act). These are one of the prerequisites for these systems to be marketed and used in the European Union. Transparency criteria can strengthen the autonomy of the stakeholders of an AI system by making informed decisions possible. Therefore, transparency can and should be considered from the outset (transparency by design).

References

- 1 ISO/IEC 22989:2022. Information technology-artificial intelligence-artificial intelligence concepts and terminology, July 2022.
- 2 BSI. Ai cloud service compliance criteria catalogue (aic4), 2021. URL: <https://www.bsi.bund.de>.
- 3 BSI. Safe, robust and comprehensible use of ai - problems, measures and needs for action, 2021. URL: <https://www.bsi.bund.de>.
- 4 Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. What's documented in ai? systematic analysis of 32k ai model cards. *CoRR*, February 2024. URL: <http://arxiv.org/abs/2402.05160>, doi: 10.48550/arXiv.2402.05160.
- 5 OECD. Recommendation of the council on artificial intelligence, 2019. URL: <https://legalinstruments.oecd.org>.

A View on Vulnerabilities: The Security Challenges of XAI

Elisabeth Pachl ✉ 

Fraunhofer Institute for Cognitive Systems, Munich, Germany

Fabian Langer ✉

TÜV Informationstechnik GmbH, Artificial Intelligence, Essen, Germany

Thora Markert ✉

TÜV Informationstechnik GmbH, Artificial Intelligence, Essen, Germany

Jeanette Miriam Lorenz ✉ 

Fraunhofer Institute for Cognitive Systems, Munich, Germany

Abstract

Modern deep learning methods have long been considered as black-boxes due to their opaque decision-making processes. Explainable Artificial Intelligence (XAI), however, has turned the tables: it provides insight into how these models work, promoting transparency that is crucial for accountability. Yet, recent developments in adversarial machine learning have highlighted vulnerabilities in XAI methods, raising concerns about security, reliability and trustworthiness, particularly in sensitive areas like healthcare and autonomous systems. Awareness of the potential risks associated with XAI is needed as its adoption increases, driven in part by the need to enhance compliance to regulations. This survey provides a holistic perspective on the security and safety landscape surrounding XAI, categorizing research on adversarial attacks against XAI and the misuse of explainability to enhance attacks on AI systems, such as evasion and privacy breaches. Our contribution includes identifying current insecurities in XAI and outlining future research directions in adversarial XAI. This work serves as an accessible foundation and outlook to recognize potential research gaps and define future directions. It identifies data modalities, such as time-series or graph data, and XAI methods that have not been extensively investigated for vulnerabilities in current research.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Explainability, XAI, Transparency, Adversarial Machine Learning, Security, Vulnerabilities

Digital Object Identifier 10.4230/OASICS.SAIA.2024.12

Category Academic Track

Funding This research was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy, as part of a project to support the thematic development of the Fraunhofer Institute for Cognitive Systems.

1 Introduction

Ever since the wide adoption of machine learning (ML), the scientific community has striven for ways to make decision-making processes based on artificial intelligence (AI) transparent, creating the field of explainable AI (XAI) [92, 46]. Transparency is critical for maintaining accountability, especially in high-risk scenarios like autonomous vehicles encountering obstacles or medical AI systems determining patient treatments. Stakeholders – including professionals utilizing AI (e.g., physicians), end-users affected by AI decisions (e.g., patients), and AI developers – benefit from understanding AI-based decisions.



© Elisabeth Pachl, Fabian Langer, Thora Markert, and Jeanette Miriam Lorenz;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 12; pp. 12:1–12:23

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The increased adoption of XAI methods is driven in part by the need to enhance compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR) [82] and the EU AI Act [83]. The GDPR preserves the “*right to explanation*” [54], encouraging organizations to provide understandable reasoning behind AI-driven decisions related to personal data. Similarly, the EU AI Act mandates transparency and human oversight for high-risk AI systems. Beyond compliance to regulations, XAI fosters trust among stakeholders, enhances the detection of biases or inaccuracies, and aligns with ethical principles such as fairness and accountability. Additionally, XAI aids continuous improvement by enabling developers to refine models based on insights from comprehensible decision-making processes.

Despite these advantages, adversarial ML (AdvML) [56, 89, 70] has become increasingly prevalent in XAI research, raising concerns regarding trustworthiness, robustness, and security [80]. This trend underscores the importance of scrutinizing XAI for potential vulnerabilities that adversarial attacks might exploit. While XAI aims to make AI systems more transparent and fair, the misuse of explainability can paradoxically be harnessed to amplify attacks on AI systems, posing threats such as evasion and privacy breaches.

Our study makes several significant contributions. We present a systematic categorization of the XAI attack surface, drawing insights from a comprehensive review of over 70 publications. This categorization helps in understanding and addressing the vulnerabilities inherent in the application of XAI. By providing concrete examples from scientific literature, we highlight specific risks associated with XAI usage. The categorization organizes attacks according to classes of XAI methods and their application domains, enabling readers to identify relevant attack vectors quickly. We discuss several aspects for the secure and robust development of AI systems incorporating XAI along the AI lifecycle. These considerations shall support the mitigation of the introduced vulnerabilities of XAI. Lastly, our work serves as a foundation for recognizing potential research gaps and defining future directions. It identifies domains and XAI methods scrutinized for vulnerabilities, guiding further investigation into countermeasures against documented attacks. It also highlights areas with scarce published attacks, suggesting potential novel attack vectors for exploration.

To the best of our knowledge, the provided information reflects the status up to March 2024. We incorporated papers from surveys on the robustness and reliability of XAI against attacks [15, 23, 77] and included notable papers from major ML conferences and journals, leveraging their citation networks to identify other relevant works.

The rest of the paper is organized as follows: Section 2 describes the background and positioning of our work within existing surveys on XAI, privacy and AdvML. Section 3 presents an overview of the attack landscape surrounding XAI, focusing on attacks on XAI and XAI-enhanced attacks. We discuss how our work can identify new attack vectors and research gaps, providing practical insights for different stakeholders of AI systems. To mitigate vulnerabilities of XAI methods, we examine certain aspects connected to the secure development of AI systems utilizing XAI in Section 4. Finally, Section 5 concludes the paper with an outlook.

2 Background

Here, we provide a brief introduction to the methodology in AdvML and XAI. Readers familiar with basic concepts of these can skip to Section 3.

2.1 Notation

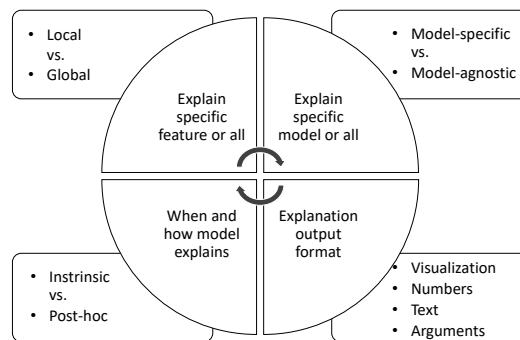
Based on Baniecki and Biecek [15], we adopt a simplified notation for our work. We mainly focus on supervised classification tasks where a model $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, parameterized by θ , maps a d -dimensional input \mathbf{x} from the feature space $\mathcal{X} \in \mathbb{R}$ to probability scores for each possible class $c \in [C]$ as a C -dimensional vector in $\mathcal{Y} \in [0, 1]^C$. The predicted class is determined by selecting the class index with the highest probability. For simplicity, we will refer to the prediction model as f . Let $\mathbf{x} \in \mathcal{X}$ represent the input vector for which we seek to explain the prediction $f(\mathbf{x})$. Consider an explanation function $g(\cdot, \cdot)$, where both the model f and \mathbf{x} serve as inputs, yielding varying outputs depending on the underlying XAI method. To facilitate the categorization of the attack surface, we use specific symbols: \rightarrow denotes a change in the given object, e.g., a small perturbation in the input: $\mathbf{x} \rightarrow \mathbf{x}'$; \approx and \neq denote similarity between two values, e.g., similar predictions $f(\mathbf{x}) \approx f(\mathbf{x}')$ or input features $\mathbf{x} \approx \mathbf{x}'$ and dissimilarity, e.g., different explanations $g(f, \mathbf{x}) \neq g(f, \mathbf{x}')$, respectively.

2.2 Explainable Artificial Intelligence

Similar to ML, XAI is a particularly wide field of research. Thus, in this section, we step back to detail the scope considered in our work. We emphasize that we do not attempt to summarize the field of XAI and refer the reader to surveys on the topic [3, 21, 26, 40, 115, 91].

XAI has found utility across various domains, including regulatory audits [57], cybersecurity [89], drug discovery [52] or model debugging [47].

Different XAI methods can be categorized based on different perspectives (Figure 1). Firstly, we distinguish between intrinsic explainable ML models and analyzing the model's outcome after training (post-hoc XAI methods). Intrinsic explainable ML models, like decision trees or attention-based neural networks, generate explanations concurrently with predictions [10]. In contrast, post-hoc explanations involve XAI methods applied at inference (e.g., Local Interpretable Model-agnostic Explanations (LIME) [87] or Gradient-weighted Class Activation Mapping (Grad-CAM) [93]). Secondly, XAI techniques can be classified as model-specific or model-agnostic. Model-specific methods are tailored to explain one specific model or a model group, while model-agnostic approaches can be applied to any ML model. The latter analyze feature importance without accessing internal model information such as weights. Examples include LIME [87], Shapley Additive Explanations (SHAP) [67], Saliency Map [51], Grad-CAM [93] or counterfactual explanations [105]. Local explainability focuses on why a specific decision was made for a single prediction instance. In contrast, global explainability offers insights into the overall decision-making process for the entire dataset. Local post-hoc feature attribution, can be obtained using perturbations-based XAI methods like Shapely values [67]. Specifically for tabular data, these allow to explain individual predictions in a model-agnostic manner. Contrary, gradient-based [68] and propagation-based [11] local post-hoc XAI methods, summarized as backpropagation-based methods in this work, e.g., Grad-CAM [93] or saliency maps [51], are specific to neural networks. Those methods leverage the principles of gradient descent to attribute importance to input features, necessitating access to the internals of the model f . Counterfactual examples are a popular approach that shows how much a specific input feature needs to change to alter the prediction outcome. Complementary to local explanations, global explanations summarize consistent patterns in model predictions across the data, such as feature importance and feature effect visualizations (e.g., partial dependence plots). For deep neural networks, concept-based explanations [43] relate human-understandable concepts to predicted classes, such as how a “stop sign” prediction is influenced by the presence of an octagon shape in an image.



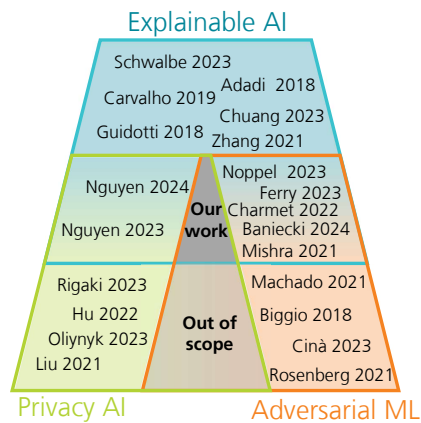
■ **Figure 1** An overview diagram showing the categorization of XAI in different aspects. Adopted from Zhang et al. [116].

In this study, we focus on post-hoc explanation methods, which offer the advantage of being versatile and applicable to a wide range of models. Additionally, the separation between the learning task and explaining its outcomes allows for evaluating threat models for the XAI method independently of the learning task. However adversaries could exploit this disparity between model’s inference and explanation, leading to discrepancies between reported predictions and explanations.

2.3 Adversarial Machine Learning

AdvML has emerged over the last 20 years as a critical research area. One major goal of advML is to – unnoticeably – alter the model’s behaviour. The most explored class of attacks focuses on the vulnerabilities of ML models to malicious inputs, known as adversarial examples [16, 31]. These adversarial inputs are strategically crafted with the intent to fool a model into misclassification, thereby posing significant threats to the reliability and trustworthiness of AI systems deployed in real-world. Adversarial examples can manifest across various data modalities, including text, tabular data, and images. For instance, in text data [7], adversaries may manipulate input text to introduce subtle changes like misspellings, and swapping words or characters [90]. Similarly, in tabular data [16, 12], adversaries may tamper with specific features to alter model predictions, whereas in image data, adversarial patches or pixels, unrecognizable by humans, added to input images can cause models to misclassify them [19, 102]. Various techniques have been developed to generate adversarial examples effectively. These techniques include gradient-based methods such as the Fast Gradient Sign Method (FGSM) [37], iterative approaches like Projected Gradient Descent (PGD) [71], and optimization-based methods like the Carlini & Wagner (C&W) attack for images [20]. Possible defenses include augmenting training data with diverse examples, model regularization [39], such as dropout and weight decay, and distillation [81], which involves training a more robust “teacher” model on original data and using its predictions as soft labels to train a “student” model.

In addition to adversarial examples, adversaries can exploit other attack vectors to compromise ML models. Backdoor attacks involve injecting malicious triggers or patterns into training data, leading to targeted misclassifications during inference [25]. These attacks typically involve poisoning the training data to ensure that the adversarial model remains indistinguishable from the desired one. Moreover, various poisoning attacks have been proposed targeting different adversarial goals, including decreasing classification accuracy or causing targeted misclassifications to evade detection. We refer to [27], for a comprehensive systematization of poisoning attacks and defenses related to model predictions.



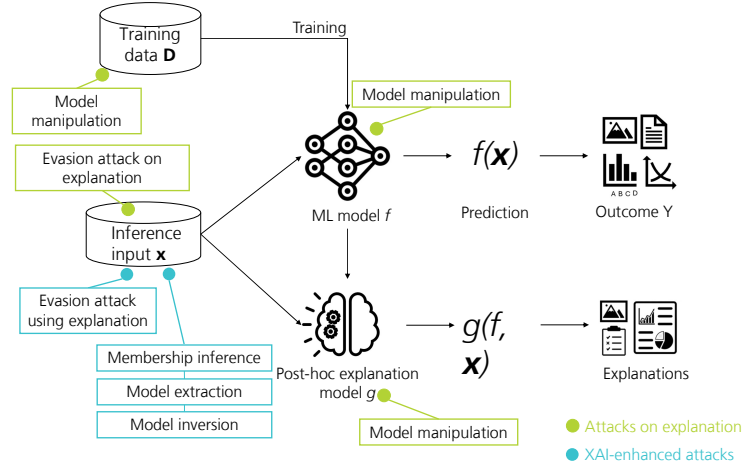
■ **Figure 2** Overview of our holistic approach combining XAI, privacy breaches, and AdvML compared to existing surveys.

Privacy attacks and model stealing attacks pose additional threats, compromising the integrity and security of AI systems. Privacy attacks seek to extract sensitive information from models [80], while model stealing attacks involve reverse-engineering model architectures or parameters using predictions or access to black-box APIs [78]. For a detailed overview of privacy attacks and defense strategies, readers are directed to the work of Rigaki et al. [88].

2.4 Comparison to Existing Surveys

Many surveys have categorized different XAI methods for ML methods and provided guidance in selecting suitable techniques for desired explanations, e.g., [3, 21, 26, 40, 115, 91], while others summarized issues related to model and data privacy in AI systems, e.g., [77, 33, 23, 15, 73] or reviewed problems related to AdvML, e.g., [70, 16, 27, 89].

Some existing surveys cover the intersection between two of the aforementioned categories (Figure 2). On the intersection between XAI and AdvML, Ferry et al. [33] examined the interplay between interpretability, fairness, and robustness, whereas Baniecki et al. [15] surveyed adversarial attacks on model explanations and fairness metrics, offering a unified taxonomy for clarity across related research areas and discussing defenses against such attacks. Noppel et al. [77] summarized attacks designed to subvert explanations based on their objectives, e.g., preserving or altering explanations, formalized notions of adversarial robustness in the presence of explanation-aware attacks, and presented a taxonomy of existing defenses. Charmet et al. [23] focused on adversarial attacks targeting XAI methods within the cybersecurity domain. They explored various attack vectors and proposed defensive strategies to maintain the fairness and integrity of XAI models. Mishra et al. [73] focused on the robustness of XAI and attacks against it. They unify existing definitions of robustness of XAI, introduce a taxonomy to classify different robustness approaches as well as some pointers about extending current robustness analysis approaches so as to identify reliable XAI methods. To the best of our knowledge, Nguyen et al. [75, 74] are the first to summarize in-depth the knowledge in the intersection between XAI and privacy AI. However, these papers only present partial coverage of the entire safety and security landscape surrounding XAI. We note, that there are works on the intersection between privacy AI and AdvML, but this is not the focus of this work.



■ **Figure 3** Attack surface against XAI systems. Trained model f predicts class label y for input x . g represents a post-hoc XAI method deriving an explanation of the input sample. Attacks can be directly against explanations (green) or XAI knowledge can be used to enhance privacy attacks or attacks against predictions (blue).

Our contribution

Our survey presents an in-depth examination of evasion and privacy breaches related to XAI, diverging from previous work by its comprehensive nature and addressing the full spectrum of possible attack vectors. We delve into the underlying principles, methodologies, and taxonomies, while also mapping out potential trajectories for future research. Especially, our work goes beyond the individual matching of defenses to specific attacks, as seen in previous studies [15, 23]. Instead, we comprehensively discussed several aspects for the secure and robust development of AI systems incorporating explicit XAI considerations along the AI lifecycle. These considerations shall support the mitigation of the introduced vulnerabilities and evolving threats of XAI.

3 Attack Landscape Surrounding XAI

While prior efforts focused predominantly on the robustness and reliability of XAI [73, 15, 77], attacks on predictions [70, 27] or XAI in AdvML [66], our work distinguishes itself by also focusing on the misuse of explainability to amplify attacks on AI systems. Broadly, the attack surface can be categorized into *attacks on XAI* and *XAI-enhanced attacks* (Figure 3). The robustness of post-hoc XAI methods and their vulnerability to adversarial examples is addressed in the context of attacks on explanations. On the other hand, when XAI is used to enhance attacks on AI systems, such as altering model predictions or compromising model privacy, these attacks fall into the category of XAI-enhanced attacks.

Table 1 and 2 lists attacks across both categories, specifying the application domain e.g., computer vision using image data, and the type of attacked XAI e.g., local vs. global or backpropagation-based vs. perturbation-based.

3.1 Attacks on XAI

We proceed to specify different types of attacks that alter explanations. We use the terms *evasion attack* and *model manipulation* based on the attack point.

3.1.1 Evasion Attack on XAI

In this scenario, an adversarial example, based on a benign inference input, is crafted to manipulate the explanation without impacting the prediction of a deployed AI system:

$$\mathbf{x} \rightarrow \mathbf{x}' \implies \begin{cases} g(f, \mathbf{x}) \neq g(f, \mathbf{x}') \\ f(\mathbf{x}) \approx f(\mathbf{x}') \end{cases}$$

Here, the adversarial example \mathbf{x}' is constructed so that its explanation matches a target explanation, while maintaining the prediction model's f output [31, 35]. Note that the target explanation differs from the original explanation of \mathbf{x} . For instance, in medical imaging, an attacker could alter an AI-interpreted CT image, changing the highlighted region indicative of malignancy or benignancy of cancer while preserving the diagnosis. This could mislead a radiologist in selecting biopsy locations, potentially compromising patient care and outcomes.

3.1.2 Model Manipulation on XAI

In this scenario, the attack involves manipulating the prediction model f or the local post-hoc explanation model g . In model manipulation of f , such as weight manipulation, fine-tuning through an expanded loss function [30, 44] or poisoning of training data [113], the altered model f' generates different explanations for the same input data \mathbf{x} , while maintaining similar predictions (e.g., [44, 30, 76]):

$$f \rightarrow f' \implies \begin{cases} g(f, \mathbf{x}) \neq g(f', \mathbf{x}) \\ f(\mathbf{x}) \approx f'(\mathbf{x}) \end{cases}$$

Further, neural networks can have backdoors triggered by specific input patterns to retrieve original explanations [104, 76]. A few works also consider how original or manipulated explanations can be used to cover an adversarial change in the model's prediction such as misclassifications [76]. This can be used to, e.g., disguise fraudulent activities in a financial fraud detection system. For instance, in a financial fraud detection system using LIME or SHAP, an attacker could manipulate the neural network's weights to disguise fraudulent transactions as legitimate. This manipulation alters the explanations to make fraudulent activity appear normal, justifying decisions that label fraudulent transactions as legitimate.

Similarly, by manipulating the local post-hoc explanation model g , the altered model g' produces different explanations for the same input data x , despite identical predictions by f [60]: $g \rightarrow g' \implies g(f, \mathbf{x}) \neq g'(f, \mathbf{x})$

While the majority of attacks are on local XAI methods, a few specifically target global XAI [60, 13, 18, 14, 62]: $g \rightarrow g' \implies \forall x \in X g(f, \mathbf{x}) \neq g'(f, \mathbf{x})$.

3.1.3 Observation

The majority of proposed attacks on explanations assume prior knowledge about the model's architecture, weights, and the XAI method used. For evasion attacks, attackers need access to the model's parameters to craft adversarial examples effectively, typically using gradient-based methods to modify inputs, changing explanations without altering predictions [44, 35]. Such attacks can be executed by individuals with significant technical expertise, such as AI developers or malicious insiders with model access. The same level of knowledge and access is necessary for model manipulation attacks, involving altering the model, XAI method, or leaving a backdoor in the model. Although generally less technically equipped, sophisticated end users with malicious intent might exploit available tools and methods to perform attacks if they can gain sufficient access to the model [29].

When considering the implementation of XAI methods, it is crucial to evaluate the advantages and disadvantages of local and global approaches. Local explanations provide insights into individual predictions, useful for case-by-case assessments, but are susceptible to adversarial manipulation, leading to significant global impacts on model behavior [44, 30, 76, 9]. Global explanations offer a comprehensive view of the model’s decision-making process but can also be manipulated to affect local explanations [60, 13, 59]. Improving detection mechanisms for one type of explanation could enhance detectability across the board [55, 35].

Generally, research on attacks aiming to alter only the explanation while preserving the prediction, such as *XAI-washing* or *fair-washing*, is sparse. We found most works studying perturbation- and permutation-based XAI methods, whereas only few exist on concept-based explanations [18], counterfactual explanations [100], and interpretable models like decision trees [62]. Additionally, explanations for language models based on text [98, 22, 50], graphs [63], and time-series data like audio [45] remain underexplored. Understanding the robustness of post-hoc XAI models to adversarial attacks in real-world applications based on underexplored data modalities is crucial. In healthcare, text data from patient records, graph data from molecular structures, and time-series data from patient monitoring systems are commonly used. In finance, transaction data, customer feedback, and network analysis for fraud detection are critical.

3.2 XAI-enhanced Attacks on Predictions

XAI methods can be exploited by adversaries to enhance attacks on AI systems. In XAI-enhanced attacks on predictions, adversarial examples are crafted using additional knowledge from XAI methods to fool AI models into making inaccurate predictions while maintaining similar explanations:

$$\mathbf{x} \rightarrow \mathbf{x}' \implies \begin{cases} g(f, \mathbf{x}) \approx g(f, \mathbf{x}') \vee g(f, \mathbf{x}) \neq g(f, \mathbf{x}') \\ f(\mathbf{x}) \neq f(\mathbf{x}') \end{cases}$$

Here, the primary goal is to make the model produce wrong predictions with consistent explanations, unlike evasion attacks where the focus is solely on altering explanations. Attackers may also aim to change both predictions and explanations to fully disguise the AI system [58], though this is more detectable due to changes in the model’s behavior.

For XAI-enhanced adversarial example crafting, gradient-based or perturbation-based explanations are used for images or tabular data to identify important pixels or features. Perturbations are added only to these areas to deceive the classifier, maintaining a high attack success rate with fewer pixel changes and reducing the optimization space and redundancy of local perturbations [41, 53, 64, 114, 2, 65]. This makes these attacks more efficient and less resource-intensive. Furthermore, with additional knowledge from XAI methods, XAI-enhanced attacks are also possible in black-box settings without any knowledge of the target model and its coupled interpreter [2, 58, 112] in contrast to white-box settings, where the attacker has full knowledge of the model [1, 53, 64]. Abdukhamidov et al. [2] demonstrated a transfer-based and score-based technique using a microbial genetic algorithm, achieving high attack success with minimal queries and high similarity in interpretations between adversarial and benign samples.

Observation

XAI-enhanced attacks on predictions pose a critical concern for AI providers due to their increased feasibility and lower execution barriers compared to attacks on explanations. We found studies across different data modalities, including image [41, 53, 64, 65, 114, 112, 1,

2, 8, 42], tabular [58], textual [107, 22], and graph data [24, 63, 108]. However, time-series data, including audio in natural language processing or sensor data from vehicle systems or patient monitoring in ICUs, remain underexplored. Overall, the potential for using XAI knowledge to enhance evasion attacks is promising but largely untapped. Although there is a substantial body of literature on the subject of evasion attacks, including work on training data poisoning, backdoor attacks and adversarial example crafting, our focus remains on XAI-enhanced attacks. A review of the literature revealed no previous studies in this field, indicating that this represents an as yet unidentified threat.

3.3 XAI-enhanced Attacks on Privacy

With the growing use of XAI methods, new vectors for privacy breaches in AI systems emerge. This section covers three primary categories of XAI-enhanced attacks on privacy: *model inversion*, *model extraction*, and *membership inferences*. These attacks leverage the additional information provided by explanations to enhance their efficiency and effectiveness.

3.3.1 Model Inversion

Model inversion attacks aim to reconstruct input data from model outputs, potentially revealing sensitive information about individuals in the training set. For example, a gender recognition scenario, an attacker might use XAI-enhanced model inversion to reconstruct facial images from the outputs of a gender classification model (prediction and explanations), leading to unauthorized re-identification and privacy violations. These attacks typically assume a black-box scenario with query-access only, where the attacker receives model predictions and explanations for a given instance \mathbf{x} . Studies have shown that explanations from backpropagation-based methods (e.g., Gradient, Grad-CAM) can significantly improve reconstruction accuracy compared to using predictions alone [117, 32, 69].

Zhao et al. [117] demonstrated enhanced model inversion attacks using XAI-aware model inversion architectures, such as multi-modal, spatially-aware CNNs. They found vulnerability varies by explanation method as they provide different levels of additional information: LRP < Gradient < CAM < Gradient x Input. Duddu et al. [32] showed that sensitive attributes can be inferred from model explanations and predictions, even when not explicitly included in input or outcome. Attacks were more successful using backpropagation-based explanations like SmoothGrad or IntegratedGradients compared to predictions alone. Luo et al. [69] focused on feature inference attacks using Shapley value, demonstrating significant advantages over prediction-only attacks in reconstructing private model inputs.

3.3.2 Model Extraction

Model extraction attacks aim to steal the functionality of a ML model by creating a surrogate model that mimics the target model's decision behaviour. Typically, the target model is deployed through an API, providing the attacker with black-box access. These attacks involve three steps: (1) collecting or synthesizing an initial unlabeled dataset, (2) querying the target model with inputs, and (3) training a surrogate model using the attack dataset annotated by the target model. It is often assumed that the attacker has an auxiliary dataset following the same distribution as the target model's training data. XAI-enhanced model extraction attacks additionally leverage explanations of queried instances. For example, in the financial sector, an attacker could use XAI-enhanced model extraction to steal a proprietary credit scoring model, undermining the original model owner's competitive advantage by replicating the sophisticated decision-making process.

Milli et al. [72] demonstrated that gradient-based explanations reveal model information more efficiently than traditional label-only queries. Their experiments showed that achieving 95% accuracy required only 10 gradient-queries, receiving predictions and explanations, compared to 1000 label-only queries for a convolutional model (CNN) on MNIST. Yan et al. [111] introduced XAMEA, an explanation-guided model extraction attack, achieving a 25% reduction in required queries for CIFAR-10 compared to traditional methods. They found Grad-CAM explanations posed the greatest risk of privacy leakage. Aïvodji et al. [5] focused on model extraction attacks using counterfactual explanations, showing an attacker could achieve over 90% fidelity with only 250 queries, significantly outperforming baseline attacks without explanations. Their findings underscored that counterfactual explanations enable high-fidelity and high-accuracy model extractions even under limited query budgets. Wang et al. [106] proposed DualCF, a querying strategy that greatly reduces the number of required queries for model extraction. Their method sequentially queries the target model with counterfactual explanations, achieving better agreement scores and lower sensitivity to sampling procedures compared to baseline methods.

3.3.3 Membership Inference

Membership inference attacks (MIAs) pose a significant threat by allowing adversaries to determine if specific data points were part of a model's training set. This can be particularly problematic in critical areas like healthcare, where sensitive information could be inferred without consent. For instance, an adversary could query a hospital's ML model for rare disease diagnosis and determine whether specific individuals' medical records were used in training.

MIAs typically assume a black-box scenario with access to model predictions and explanations. They aim to predict the membership status of data points within the attack set. XAI-enhanced and prediction-only MIAs use two main strategies: threshold-based and reference model-based attacks [96]. Threshold-based attacks rely on output variance, assuming training set data points yield lower variance in predictions and explanations due to the model's familiarity with the data. Reference model-based attacks use shadow models to simulate the target model's behavior and derive membership inference thresholds. This method assumes access to similar data and knowledge of the model's architecture and hyperparameters.

Shokri et al. [96] pioneered investigating using model explanations for inferring private information about training data. They proposed a threshold-based attack using prediction and explanation variance, revealing significant privacy risks with backpropagation-based explanations in tabular datasets, but not in image datasets. They attributed this to fluctuating gradient variance. While it is entirely possible that perturbation-based methods are vulnerable to membership inference, the authors conjecture that this is not the case. Pawelczyk et al. [84] highlighted privacy risks from algorithmic recourse, introducing counterfactual distance-based attacks that infer membership without auxiliary data or model details. These attacks excelled with overfitting models and high data dimensionality. Goethals et al. [36] introduced explanation linkage attacks, where adversaries use quasi-identifiers from counterfactual explanations to re-identify individuals by linking with background information. They also proposed k-anonymous counterfactual explanations to mitigate these risks.

3.3.4 Observation

XAI-enhanced privacy attacks significantly increase the effectiveness and efficiency of model and data privacy breaches, making them more feasible in real-world scenarios. These attacks require minimal prior knowledge and model access and reduce the number of queries needed

for successful breaches. The effectiveness of MIAs varies by data modality, with tabular and high-dimensional data being more susceptible [96]. Additionally, backpropagation-based methods are more vulnerable to MIAs than perturbation-based methods [96]. Counterfactual explanations pose significant risks for both MIAs and model extraction attacks [5, 106, 84, 36], although no work has been found addressing model inversion using counterfactual explanations. Research gaps exist in studying XAI methods' vulnerability in MIAs for tabular data and model inversion attacks using counterfactual explanations. Current model inversion studies focus primarily on tabular data [32, 69], with no work on textual data.

3.4 Practical Applications of the XAI Attack Vector Classification Table

The application of XAI methods can have a significant impact on a system's security and safety. Depending on the system at hand and the implemented XAI method, different attack vectors may apply. Tables 1 and 2 provide an overview of published attacks against XAI and XAI-enhanced attacks against AI systems, extending Baniecki et al.'s work [15].

The tables arrange studies by data modalities, groups of XAI methods, and attack types (privacy, prediction, and attacks on XAI). More granular attack subcategories further describe the type of attack presented in the referenced works. Table 1 covers the topic of computer vision, while the papers introduced in Table 2 deal with graphs, textual, numerical and time-series/audio data.

These tables serve multiple stakeholders: *Developers* can identify potential vulnerabilities early in the design and development stage. By understanding specific attack vectors associated with different XAI methods, they can proactively implement countermeasures and design more secure models. Section 4 shall support the development of secure and safe AI systems leveraging XAI. *Users* gain insights into limitations and risks associated with system explanations, recognizing potential compromises or errors [17]. An overview of attacks based on explainability methods equips *evaluators* with the necessary knowledge to conduct thorough and informed risk assessments and later on perform targeted vulnerability testing to verify the system's robustness against these kind of attacks. *Researchers* can identify knowledge gaps, explore new attack vectors, develop novel defense mechanisms, and enhance existing XAI methods.

4 Aspects of XAI Attack Mitigation

Our comprehensive analysis of potential attacks on and enhanced by XAI should not deter its use but rather highlight latent risks. Despite these risks, however, explainability offers such significant benefits that it should not be dispensed with. Under certain circumstances, it may even be necessary to use XAI methods in order to improve adherence with transparency obligations, such as those stated in the EU AI Act [83].

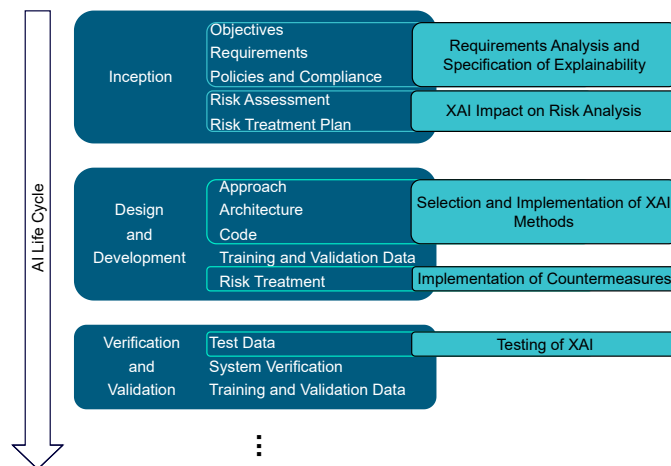
In the following, we present aspects connected to the responsible implementation and use of XAI methods throughout the first phases of the AI life cycle in accordance with ISO/IEC 22989 [49]. For the phases from inception to verification and validation, specific considerations are highlighted in order to mitigate potential risks and ensure the secure and safe use of XAI (Figure 4).

■ **Table 1** Summary of adversarial attacks on explanations and XAI-enhanced attacks on model predictions and privacy for data modality of images.

Data Modality	XAI Method	XAI-enhanced on Privacy			XAI-enhanced Attacks on Predictions	Attacks on XAI	
		Model Inversion	Model Extraction	Membership Inference		Model Manipulation	Evasion Attack
Computer Vision	Local	Zhao [117]	Yan [111], Milli [72], Yan [110], Yan [109]	Shokri [96]	Guo [41], Jing [53], Liu [64], Zhan [112], Zhang [114], Abdukhamidov [1], Abdukhamidov [2]	Heo [44], Nopel [76], Kindermans [55], Viering [104], Aïvodji [4], Zhang [113]	Anders [9], De Aguiar [29], Dombrowski [31], Ghorbani [35], Göpfert [38], Galli [34], Huang [48], Le [61], Pandya [79], Rasae [85], Renkhoff [86], Song [101], Subramanya [103], Zhang [114], Kindermans [55], Si [97]
	Backpropagation		Yan [111], Yan [110], Yan [109]	Shokri [96]	Amich [8], Liu [65]		Göpfert [38], Pandya [79]
					Hada [42]		
	Concept-based					Brown [18]	

Table 2 Summary of adversarial attacks on explanations and XAI-enhanced attacks on model predictions and privacy for data modality of textual, numerical, graph and time-series (TS) including audio data.

Data Modality	XAI Method	XAI-enhanced on Privacy				XAI-enhanced Attacks on Predictions		Attacks on XAI	
		Model Inversion	Model Extraction	Membership Inference	Evasion Attack	Model Manipulation	Evasion Attack	Model Manipulation	Evasion Attack
Textual & Numerical	Local	Backpropagation	Duddu [32]			Kuppa and Le-Khac [58], Xu and Du [107]		Dimanov [30], Zhang [113]	Ali [6], Anders [9], Ivankay [50], Sinha [98], Kuppa and Le-Khac [58]
		Perturbation	Duddu [32]			Chai [22]		Baniecki and Biecek [13], Dimanov [30], Severi [94]	Ali [6], Sinha [98], Slack [99]
		Counterfactual			Aivodji [5], Wang [106]	Pawelczyk [84], Goethals [36]		Slack [100]	
	Global	Backpropagation	Luo [69]						
		Perturbation						Baniecki and Biecek [13], Baniecki [14]	Laberge [59], Lakkaraju [60]
		Interpretable Models						Le Merrer [62]	
	Local	Backpropagation + perturbation				Chen [24], Li [63], Xu [108]			Li [63]
	Audio/TS	Backpropagation							Hoedt [45]
		Perturbation							Hoedt [45]



■ **Figure 4** Aspects for the secure development of AI systems incorporating XAI along the AI life cycle.

Requirements Analysis and Specification of Explainability

In inception, assessing the necessity and benefits of explainability is crucial. The superior reason for integrating methods for explainability into a system is creating transparency for different stakeholders, providing insight into the system's general functionality or specific model operations. It is an important step to be clear in advance about the requirements coming from various sides that need to be fulfilled. For example, these requirements may originate from regulation (e.g., Article 13 of the EU AI Act [83]), adaption of standards and best-practices (e.g., Microsoft's Responsible AI Standard [28]) or business goals. Subsequently, the identified requirements must be specified regarding use case (including used data), the planned system architecture and environment. The goal is to formulate concise requirements for explainability, so that only the required level of transparency is provided and no unnecessary information is disclosed.

XAI Impact on Risk Analysis

The integration of XAI into a system can introduce new risks and potential attack vectors, significantly affecting risk analysis. While XAI enhances transparency and trust in AI systems by providing clear and interpretable insights, it also necessitates a thorough reassessment of security vulnerabilities and risk management strategies. One primary risk introduced by XAI is the potential exposure of the model's inner workings to adversaries. XAI methods reveal how models make decisions, inadvertently disclosing sensitive aspects like feature importance and decision pathways. This transparency can be exploited to launch targeted attacks (Section 3.2). Thus, detailed insights provided by XAI necessitate robust security measures to protect the model from exploitation. Additionally, XAI techniques can increase the risk of privacy attacks (Section 3.3). This is particularly concerning in applications involving personal or confidential data, such as healthcare or financial services. The enhanced interpretability offered by XAI can make it easier for attackers to infer training data and thus private information. Corresponding privacy-preserving techniques shall be considered. The integrity of the explanations themselves is another critical concern. If explanations can be manipulated (Section 3.1), the trustworthiness of the entire AI system can be compromised. Attackers might alter explanations to hide malicious activities or to falsely assure users of

the model's reliability. The potential attack vectors depend on the selected XAI method and the domain the system is operating in. Table 1 and Table 2 provide a state-of-the-art overview of potential attacks based on various XAI methods in different domains to support the risk analysis process.

Despite these challenges, integrating XAI can enhance overall risk management by providing clearer insights into model behavior and decision-making processes. This transparency can help identifying potential biases and vulnerabilities within the model, enabling more effective mitigation strategies. By understanding how models arrive at their decisions, organizations can implement targeted defenses against specific risks and continuously monitor and improve the AI system's security posture.

Selection and Implementation of XAI and Countermeasures

As mentioned, implementing XAI methods introduces further risks and attack vectors based on the information obtained by these methods. Therefore, mitigating these threats requires balancing transparency with security. Providing too much detail in explanations can expose the model to various attacks, while insufficient transparency can undermine XAI's purpose, which is to build trust and understanding. Striking the right balance involves carefully selecting suitable methods to provide necessary insights without disclosing sensitive information that could be exploited. The XAI method should be chosen strictly based on the determined requirements for the type and extent of explainability needed. The explanations themselves can be a target for tampering. Ensuring the integrity and authenticity of explanations through cryptographic techniques like digital signatures can help verify that the explanations have not been altered and are legitimate [95]. Furthermore, explainability methods can inadvertently expose sensitive aspects of the "explained" AI model, such as proprietary algorithms or business logic. Role-based access controls can ensure that only authorized personnel view detailed explanations, protecting intellectual property and sensitive information. Employing robust adversarial training techniques can also help the model resist adversarial attacks based on or enhanced by XAI.

Testing of XAI

Initial testing should focus on verifying the introduced requirements for XAI. Test cases shall ensure that only necessary information is published, but that explanations are still effective. Therefore, testing XAI has to be conducted especially from the viewpoint of the target group of the explanations. Additionally, vulnerability testing shall be conducted with regard to XAI-related attacks, e.g., as listed in Table 1 and Table 2. Research for state-of-the-art attacks should always be carried out and relevant attacks are to be incorporated in the vulnerability testing activities.

5 Conclusion

As XAI methods move from research to practical applications, concerns about malicious use and adversarial attacks have increased. This work provides a comprehensive overview of security and robustness issues in XAI, categorizing research on adversarial attacks targeting ML explanations and the exploitation of explainability to enhance attacks on AI systems. Most studies focus on predictive models using imaging and tabular datasets with backpropagation and perturbation-based XAI techniques. Further research is needed on adversarial attacks in other data modalities, such as language, graphs, time series, multimodal systems, and

explanations for reinforcement learning agents and transformer-based generative AI like large language models. Additionally, this review highlights the need to evaluate vulnerabilities in intrinsically explainable ML architectures, such as decision trees and attention-based neural networks, and how their explanations could enhance attacks.

Practically, integrating XAI into AI systems requires awareness of its dual-edged nature. While XAI offers benefits like compliance, user trust, and system debugging, it also introduces security risks that must be mitigated to ensure the safe development and deployment. Therefore, the integration of XAI into AI systems requires a thorough assessment of the potential risks and corresponding countermeasures. XAI methods should be selected carefully to ensure explanations are informative without revealing sensitive information that could facilitate attacks on the AI system.

References

- 1 E. Abdukhmidov, M. Abuhamad, F. Juraev, E. Chan-Tin, and T. AbuHmed. AdvEdge: Optimizing Adversarial Perturbations Against Interpretable Deep Learning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13116 LNCS, pages 93–105, 2021. doi:10.1007/978-3-030-91434-9_9.
- 2 E. Abdukhmidov, F. Juraev, M. Abuhamad, and T. Abuhmed. Black-box and Target-specific Attack Against Interpretable Deep Learning Systems. In *ASIA CCS 2022 - Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security*, pages 1216–1218, 2022. doi:10.1145/3488932.3527283.
- 3 Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. doi:10.1109/ACCESS.2018.2870052.
- 4 U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: The risk of rationalization. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 240–252, 2019.
- 5 Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations, September 2020. arXiv:2009.01884, doi:10.48550/arXiv.2009.01884.
- 6 H. Ali, M.S. Khan, A. Al-Fuqaha, and J. Qadir. Tamp-X: Attacking explainable natural language classifiers through tampered activations. *Computers and Security*, 120, 2022. doi:10.1016/j.cose.2022.102791.
- 7 Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018. arXiv:1804.07998.
- 8 Abderrahmen Amich and Birhanu Eshete. EG-Booster: Explanation-Guided Booster of ML Evasion Attacks. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy, CODASPY '22*, pages 16–28, New York, NY, USA, April 2022. Association for Computing Machinery. doi:10.1145/3508398.3511510.
- 9 C.J. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Muller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF168147-1, pages 291–300, 2020.
- 10 Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.
- 11 Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

- 12 Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. Imperceptible adversarial attacks on tabular data. *arXiv preprint arXiv:1911.03274*, 2019. [arXiv:1911.03274](#).
- 13 H. Baniecki and P. Biecek. Manipulating SHAP via Adversarial Data Perturbations (Student Abstract). In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 12907–12908, 2022.
- 14 H. Baniecki, W. Kretowicz, and P. Biecek. Fooling Partial Dependence via Data Poisoning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13715 LNAI, pages 121–136, 2023. [doi:10.1007/978-3-031-26409-2_8](#).
- 15 Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, page 102303, 2024. [doi:10.1016/J.INFFUS.2024.102303](#).
- 16 Battista Biggio, Iginio Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013. [doi:10.1007/978-3-642-40994-3_25](#).
- 17 Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Why do explanations fail? a typology and discussion on failures in xai. *arXiv preprint arXiv:2405.13474*, 2024. [doi:10.48550/arXiv.2405.13474](#).
- 18 D. Brown and H. Kvinge. Making Corgis Important for Honeycomb Classification: Adversarial Attacks on Concept-based Explainability Tools. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2023-June, pages 620–627, 2023. [doi:10.1109/CVPRW59228.2023.00069](#).
- 19 Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. [arXiv:1712.09665](#).
- 20 Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017. [doi:10.1109/SP.2017.49](#).
- 21 Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.
- 22 Y. Chai, R. Liang, S. Samtani, H. Zhu, M. Wang, Y. Liu, and Y. Jiang. Additive Feature Attribution Explainable Methods to Craft Adversarial Attacks for Text Classification and Text Regression. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2023. [doi:10.1109/TKDE.2023.3270581](#).
- 23 Fabien Charmet, Harry Chandra Tanuwidjaja, Solayman Ayoubi, Pierre-François Gimenez, Yufei Han, Houda Jmila, Gregory Blanc, Takeshi Takahashi, and Zonghua Zhang. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, 77(11):789–812, 2022. [doi:10.1007/S12243-022-00926-7](#).
- 24 L. Chen, N. Yan, B. Zhang, Z. Wang, Y. Wen, and Y. Hu. A General Backdoor Attack to Graph Neural Networks Based on Explanation Method. In *Proceedings - 2022 IEEE 21st International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2022*, pages 759–768, 2022. [doi:10.1109/TrustCom56396.2022.00107](#).
- 25 Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. [arXiv:1712.05526](#).
- 26 Yu-Neng Chuang, Guanchu Wang, Fan Yang, Zirui Liu, Xuanning Cai, Mengnan Du, and Xia Hu. Efficient xai techniques: A taxonomic survey. *arXiv preprint arXiv:2302.03225*, 2023.

- 27 Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s):1–39, 2023. doi:10.1145/3585385.
- 28 Microsoft Corporation. Microsoft responsible ai standard, v2, 2022. accessed 29 July 2024. URL: <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf>.
- 29 E.J. De Aguiar, M.V.L. Costa, C. Traina, and A.J.M. Traina. Assessing Vulnerabilities of Deep Learning Explainability in Medical Image Analysis under Adversarial Settings. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, volume 2023-June, pages 13–16, 2023. doi:10.1109/CBMS58004.2023.00184.
- 30 Boty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You Shouldn’t Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In Huáscar Espinoza, José Hernández-Orallo, Xin Cynthia Chen, Seán S. ÓhÉigeartaigh, Xiaowei Huang, Mauricio Castillo-Effen, Richard Mallah, and John A. McDermid, editors, *Proceedings of the Workshop on Artificial Intelligence Safety, Co-Located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 63–73. CEUR-WS.org, 2020. URL: <https://ceur-ws.org/Vol-2560/paper8.pdf>.
- 31 Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- 32 V. Duddu and A. Boutet. Inferring Sensitive Attributes from Model Explanations. In *International Conference on Information and Knowledge Management, Proceedings*, pages 416–425, 2022. doi:10.1145/3511808.3557362.
- 33 Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Sok: Taming the triangle—on the interplays between fairness, interpretability and privacy in machine learning. *arXiv preprint arXiv:2312.16191*, 2023. doi:10.48550/arXiv.2312.16191.
- 34 A. Galli, S. Marrone, V. Moscato, and C. Sansone. Reliability of eXplainable Artificial Intelligence in Adversarial Perturbation Scenarios. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12663 LNCS, pages 243–256, 2021. doi:10.1007/978-3-030-68796-0_18.
- 35 A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 3681–3688, 2019.
- 36 Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):83:1–83:24, August 2023. doi:10.1145/3608482.
- 37 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- 38 Jan Philip Göpfert, Heiko Wersing, and Barbara Hammer. Recovering localized adversarial attacks. In *Artificial Neural Networks and Machine Learning—ICANN 2019: Theoretical Neural Computation: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part I 28*, pages 302–311. Springer, 2019. doi:10.1007/978-3-030-30487-4_24.
- 39 Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

- 40 Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018. doi:10.1145/3236009.
- 41 S. Guo, S. Geng, T. Xiang, H. Liu, and R. Hou. ELAA: An efficient local adversarial attack using model interpreters. *International Journal of Intelligent Systems*, 37(12):10598–10620, 2022. doi:10.1002/int.22680.
- 42 S.S. Hada, M.Á. Carreira-Perpiñán, and A. Zharmagambetov. Sparse oblique decision trees: A tool to understand and manipulate neural net features. *Data Mining and Knowledge Discovery*, 2023. doi:10.1007/s10618-022-00892-7.
- 43 Lena Heidemann, Maureen Monnet, and Karsten Roscher. Concept correlation and its effects on concept-based models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4780–4788, 2023.
- 44 J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- 45 K. Hoedt, V. Praher, A. Flexer, and G. Widmer. Constructing adversarial examples to investigate the plausibility of explanations in deep audio and image classifiers. *Neural Computing and Applications*, 35(14):10011–10029, 2023. doi:10.1007/s00521-022-07918-7.
- 46 Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemysław Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pages 13–38. Springer, 2022.
- 47 Weronika Hryniewska, Przemysław Bombiński, Patryk Szatkowski, Paulina Tomaszewska, Artur Przelaskowski, and Przemysław Biecek. Checklist for responsible deep learning modeling of medical images based on covid-19 detection studies. *Pattern Recognition*, 118:108035, 2021. doi:10.1016/J.PATCOG.2021.108035.
- 48 Q. Huang, L. Chiang, M. Chiu, and H. Sun. Focus-Shifting Attack: An Adversarial Attack That Retains Saliency Map Information and Manipulates Model Explanations. *IEEE Transactions on Reliability*, pages 1–12, 2023. doi:10.1109/TR.2023.3303923.
- 49 International Standardization Organization (ISO). Iso/iec 22989:2022 artificial intelligence concepts and terminology, 2022.
- 50 A. Ivankay, I. Girardi, C. Marchiori, and P. Frossard. FOOLING EXPLANATIONS IN TEXT CLASSIFIERS. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.
- 51 Rahul Iyer, Yuezhong Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018. doi:10.1145/3278721.3278776.
- 52 José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020. doi:10.1038/S42256-020-00236-4.
- 53 H. Jing, C. Meng, X. He, and W. Wei. Black Box Explanation Guided Decision-Based Adversarial Attacks. In *2019 IEEE 5th International Conference on Computer and Communications, ICC3 2019*, pages 1592–1596, 2019. doi:10.1109/ICCC47050.2019.9064243.
- 54 Margot E Kaminski. The right to explanation, explained (june 15, 2018). university of colorado law legal studies research paper no. 18-24. *Berkeley Technology Law Journal*, 34(1), 2019.
- 55 Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 267–280. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-28954-6_14.
- 56 Zico Kolter and Aleksander Madry. Adversarial robustness: Theory and practice. *Tutorial at NeurIPS*, page 3, 2018.

- 57 Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. In *International Conference on Machine Learning*, pages 17808–17826. PMLR, 2023. URL: <https://proceedings.mlr.press/v202/krishna23a.html>.
- 58 A. Kuppa and N.-A. Le-Khac. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In *Proceedings of the International Joint Conference on Neural Networks*, 2020. doi:10.1109/IJCNN48605.2020.9206780.
- 59 Gabriel Laberge, Ulrich Aïvodji, Satoshi Hara, Mario Marchand, and Foutse Khomh. Fool SHAP with Stealthily Biased Sampling. In *International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.
- 60 H. Lakkaraju and O. Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020. doi:10.1145/3375627.3375833.
- 61 Thi-Thu-Huong Le, Hyoeun Kang, and Howon Kim. Robust Adversarial Attack Against Explainable Deep Classification Models Based on Adversarial Images With Different Patch Sizes and Perturbation Ratios. *IEEE Access*, 9:133049–133061, 2021. doi:10.1109/ACCESS.2021.3115764.
- 62 Erwan Le Merrer and Gilles Trédan. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2(9):529–539, September 2020. doi:10.1038/s42256-020-0216-z.
- 63 Yiqiao Li, Sunny Verma, Shuiqiao Yang, Jianlong Zhou, and Fang Chen. Are graph neural network explainers robust to graph noises? In *Australasian Joint Conference on Artificial Intelligence*, pages 161–174. Springer, 2022. doi:10.1007/978-3-031-22695-3_12.
- 64 Haohan Liu, Xingquan Zuo, Hai Huang, and Xing Wan. Saliency map-based local white-box adversarial attack against deep neural networks. In *CAAI International Conference on Artificial Intelligence*, pages 3–14. Springer, 2022. doi:10.1007/978-3-031-20500-2_1.
- 65 Mingting Liu, Xiaozhang Liu, Anli Yan, Yuan Qi, and Wei Li. Explanation-Guided Minimum Adversarial Attack. In Yuan Xu, Hongyang Yan, Huang Teng, Jun Cai, and Jin Li, editors, *Machine Learning for Cyber Security*, Lecture Notes in Computer Science, pages 257–270, Cham, 2023. Springer Nature Switzerland. doi:10.1007/978-3-031-20096-0_20.
- 66 Ninghao Liu, Mengnan Du, Ruocheng Guo, Huan Liu, and Xia Hu. Adversarial attacks and defenses: An interpretation perspective. *ACM SIGKDD Explorations Newsletter*, 23(1):86–99, 2021. doi:10.1145/3468507.3468519.
- 67 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- 68 Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR, 2022. URL: <https://proceedings.mlr.press/v162/lundstrom22a.html>.
- 69 X. Luo, Y. Jiang, and X. Xiao. Feature Inference Attack on Shapley Values. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 2233–2247, 2022. doi:10.1145/3548606.3560573.
- 70 Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender’s perspective. *ACM Computing Surveys (CSUR)*, 55(1):1–38, 2021.
- 71 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. arXiv:1706.06083.
- 72 Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. Model Reconstruction from Model Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, pages 1–9, New York, NY, USA, January 2019. Association for Computing Machinery. doi:10.1145/3287560.3287562.

- 73 Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzini. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021. [arXiv:2111.00358](#).
- 74 Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024. [doi:10.48550/arXiv.2404.00673](#).
- 75 Truc Nguyen, Phung Lai, Hai Phan, and My T Thai. Xrand: Differentially private defense against explanation-guided attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11873–11881, 2023. [doi:10.1609/AAAI.V37I10.26401](#).
- 76 Maximilian Noppel, Lukas Peter, and Christian Wressnegger. Disguising Attacks with Explanation-Aware Backdoors. In *2023 IEEE Symposium on Security and Privacy (SP)*, page 664, 2023. [doi:10.1109/SP46215.2023.10179308](#).
- 77 Maximilian Noppel and Christian Wressnegger. Sok: Explainable machine learning in adversarial environments. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 21–21. IEEE Computer Society, 2023.
- 78 Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55(14s):1–41, 2023. [doi:10.1145/3595292](#).
- 79 M.A. Pandya, P.C. Siddalingaswamy, and S. Singh. Explainability of Image Classifiers for Targeted Adversarial Attack. In *INDICON 2022 - 2022 IEEE 19th India Council International Conference*, 2022. [doi:10.1109/INDICON56171.2022.10039871](#).
- 80 Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*, pages 399–414. IEEE, 2018. [doi:10.1109/EUROSP.2018.00035](#).
- 81 Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. [doi:10.1109/SP.2016.41](#).
- 82 The European Parliament and The Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.
- 83 The European Parliament and The Council of the European Union. Artificial intelligence act, 2024. accessed 29 July 2024. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.
- 84 Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. On the Privacy Risks of Algorithmic Recourse. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 9680–9696. PMLR, April 2023. URL: <https://proceedings.mlr.press/v206/pawelczyk23a.html>.
- 85 H. Rasaei and H. Rivaz. Explainable AI and susceptibility to adversarial attacks: A case study in classification of breast ultrasound images. In *IEEE International Ultrasonics Symposium, IUS*, 2021. [doi:10.1109/IUS52206.2021.9593490](#).
- 86 J. Renkhoff, W. Tan, A. Velasquez, W.Y. Wang, Y. Liu, J. Wang, S. Niu, L.B. Fazlic, G. Dartmann, and H. Song. Exploring Adversarial Attacks on Neural Networks: An Explainable Approach. In *Conference Proceedings of the IEEE International Performance, Computing, and Communications Conference*, volume 2022-November, pages 41–42, 2022. [doi:10.1109/IPCCC55026.2022.9894322](#).
- 87 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- 88 Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023. [doi:10.1145/3624010](#).

- 89 Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021. doi:10.1145/3453158.
- 90 Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. Models in the wild: On corruption robustness of neural nlp systems. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26*, pages 235–247. Springer, 2019. doi:10.1007/978-3-030-36718-3_20.
- 91 Maresa Schröder, Alireza Zamanian, and Narges Ahmidi. Post-hoc saliency methods fail to capture latent feature importance in time series data. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 106–121. Springer, 2023. doi:10.1007/978-3-031-39539-0_10.
- 92 Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.
- 93 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. doi:10.1109/ICCV.2017.74.
- 94 G. Severi, J. Meyer, S. Coull, and A. Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In *Proceedings of the 30th USENIX Security Symposium*, pages 1487–1504, 2021.
- 95 Rucha Shinde, Shruti Patil, Ketan Kotecha, Vidyasagar Potdar, Ganeshsree Selvachandran, and Ajith Abraham. Securing ai-based healthcare systems using blockchain technology: A state-of-the-art systematic literature review and future research directions. *Transactions on Emerging Telecommunications Technologies*, 2024.
- 96 R. Shokri, M. Strobel, and Y. Zick. On the Privacy Risks of Model Explanations. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021. doi:10.1145/3461702.3462533.
- 97 N. Si, H. Chang, and Y. Li. A Simple and Effective Method to Defend Against Saliency Map Attack. In *ACM International Conference Proceeding Series*, 2021. doi:10.1145/3474198.3478141.
- 98 Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.blackboxnlp-1.33.
- 99 D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. doi:10.1145/3375627.3375830.
- 100 D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, volume 1, pages 62–75, 2021.
- 101 Qianqian Song, Xiangwei Kong, and Ziming Wang. Fooling Neural Network Interpretations: Adversarial Noise to Attack Images. In Lu Fang, Yiran Chen, Guangtao Zhai, Jane Wang, Ruiping Wang, and Weisheng Dong, editors, *Artificial Intelligence*, Lecture Notes in Computer Science, pages 39–51, Cham, 2021. Springer International Publishing. doi:10.1007/978-3-030-93049-3_4.
- 102 Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. doi:10.1109/TEVC.2019.2890858.

- 103 A. Subramanya, V. Pillai, and H. Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, pages 2020–2029, 2019. doi:10.1109/ICCV.2019.00211.
- 104 T. VIERING, Ziqi Wang, M. Loog, and E. Eisemann. How to Manipulate CNNs to Make Them Lie: The GradCAM Case. *ArXiv*, July 2019.
- 105 Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.
- 106 Y. Wang, H. Qian, and C. Miao. DualCF: Efficient Model Extraction Attack from Counterfactual Explanations. In *ACM International Conference Proceeding Series*, pages 1318–1329, 2022. doi:10.1145/3531146.3533188.
- 107 J. Xu and Q. Du. Adversarial attacks on text classification models using layer-wise relevance propagation. *International Journal of Intelligent Systems*, 35(9):1397–1415, 2020. doi:10.1002/int.22260.
- 108 J. Xu, M. Xue, and S. Picek. Explainability-based Backdoor Attacks against Graph Neural Networks. In *WiseML 2021 - Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*, pages 31–36, 2021. doi:10.1145/3468218.3469046.
- 109 A. Yan, R. Hou, X. Liu, H. Yan, T. Huang, and X. Wang. Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956, 2022. doi:10.1002/int.23022.
- 110 A. Yan, R. Hou, H. Yan, and X. Liu. Explanation-based data-free model extraction attacks. *World Wide Web*, 26(5):3081–3092, 2023. doi:10.1007/s11280-023-01150-6.
- 111 A. Yan, T. Huang, L. Ke, X. Liu, Q. Chen, and C. Dong. Explanation leaks: Explanation-guided model extraction attacks. *Information Sciences*, 632:269–284, 2023. doi:10.1016/j.ins.2023.03.020.
- 112 Y. Zhan, B. Zheng, Q. Wang, N. Mou, B. Guo, Q. Li, C. Shen, and C. Wang. Towards Black-Box Adversarial Attacks on Interpretable Deep Learning Systems. In *Proceedings - IEEE International Conference on Multimedia and Expo*, volume 2022-July, 2022. doi:10.1109/ICME52920.2022.9859856.
- 113 H. Zhang, J. Gao, and L. Su. Data Poisoning Attacks against Outcome Interpretations of Predictive Models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2165–2173, 2021. doi:10.1145/3447548.3467405.
- 114 X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *Proceedings of the 29th USENIX Security Symposium*, pages 1659–1676, 2020.
- 115 Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. doi:10.1109/TETCI.2021.3100641.
- 116 Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10:93104–93139, 2022. doi:10.1109/ACCESS.2022.3204051.
- 117 X. Zhao, W. Zhang, X. Xiao, and B. Lim. Exploiting Explanations for Model Inversion Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 662–672, 2021. doi:10.1109/ICCV48922.2021.00072.


AI Certification: Empirical Investigations into Possible Cul-De-Sacs and Ways Forward

Benjamin Fresz ✉ 

Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

Danilo Brajovic ✉ 

Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

Marco F. Huber ✉ 

Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Germany

Abstract

In this paper, previously conducted studies regarding the development and certification of safe Artificial Intelligence (AI) systems from the practitioner's viewpoint are summarized. Overall, both studies point towards a common theme: AI certification will mainly rely on the analysis of the processes used to create AI systems. While additional techniques such as methods from the field of eXplainable AI (XAI) and formal verification methods seem to hold a lot of promise, they can assist in creating safe AI-systems, but do not provide comprehensive solutions to the existing problems in regard to AI certification.

2012 ACM Subject Classification Social and professional topics → Testing, certification and licensing; Computing methodologies → Machine learning; General and reference → Empirical studies

Keywords and phrases AI certification, eXplainable AI (XAI), safe AI, trustworthy AI, AI documentation

Digital Object Identifier 10.4230/OASICS.SAIA.2024.13

Category Practitioner Track

Funding This paper is funded in parts by the German Federal Ministry for Economic Affairs and Climate Action under grant no. 19A21040B (project “veoPipe”), an OEM company via a joint project and by the Fraunhofer Gesellschaft under grant no. PREPARE 40-02702 (project “ML4Safety”).

Acknowledgements Parts of this paper were refined with the help of company-specific LLM (Fh-Genie 4o).

1 Introduction

Artificial Intelligence (AI) has rapidly integrated into various industries, necessitating the development of robust certification standards to ensure reliability, safety, and ethical compliance. Current challenges include the lack of standardized guidelines and the opaqueness of AI decision-making processes, which can lead to mistrust and potential misuse [4, 5, 7]. AI certification could – when done correctly – ensure the performance and trustworthiness of AI systems. Concurrently, Explainable AI (XAI) addresses the need for transparency and interpretability in AI models, aiming to make AI decisions comprehensible to humans, foster trust, and enable informed decision-making.



© Benjamin Fresz, Danilo Brajovic, and Marco F. Huber;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 13; pp. 13:1–13:4

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 The EU AI Act

The European Union's AI Act [1] proposes a regulatory framework to govern AI systems, emphasizing transparency, accountability, and risk management. The AI Act categorizes applications into different risk levels, each requiring specific compliance measures. However, the AI Act lacks technical details, which are expected to be specified in future harmonized standards. This regulatory uncertainty underscores the need for practical insights and frameworks to guide the effective certification of AI systems in real-world settings. For example, such guidelines could provide clear information on how to document the development process of AI applications or the selection process of appropriate training data. Industry practitioners and researchers can contribute valuable insights by sharing experiences and developing best practices that align with the goals of transparency, accountability, and risk management.

3 Empirical Studies

Two studies conducted in Germany aimed to gather such insights [2, 3]. The first study evaluated general opinions regarding AI safeguarding within industry, while the second focused on the use of XAI methods for certification purposes.

3.1 Effectively Documenting AI Applications

The first study reviewed activities surrounding the safeguarding and certification of AI systems over recent years. Collaborations with industry partners in Germany across various domains (automotive, finance, and food manufacturing) and interviews with certification experts, employees, and developers led to the publication of a framework for documenting AI applications along four major development steps [2]. The feedback from practitioners highlighted five main points:

The AI Assessment Catalogue [6] is widely used in Germany and regarded as a de facto standard, but its length (approximately 160 pages) poses practical challenges. It is primarily designed for final assessments, not for providing support during the development phase.

There is some *uncertainty about the Assessment Catalog*, as practitioners are concerned about its compliance with the AI Act and its extensive length, though it remains the primary tool for developing safe AI.

Partners expressed *high hopes for future standards*, as they are eagerly awaiting new standards for safe AI development, showing interest in best practices for data collection and model selection, despite there being some concern about the practical usefulness of these standards.

Need for implementation details, technical tools, and clearly defined performance thresholds was often expressed, as partners desire more concrete guidance, including specific tools and methods, actionable instructions, and specific information to facilitate the development process.

Trust-Building is a key motivation to pursue certification of AI systems. Affected individuals want to be included in the development process and understand the design choices made.

3.2 XAI for Safe AI and Certification

The second study focused on experts in XAI and AI certification to explore expectations regarding the use of XAI in this area [3]. Through 15 qualitative interviews, it was found that XAI can help to identify errors such as biases within AI models. However, it cannot comprehensively answer the question, “Is this AI model safe to use?”

Surrounding the use of XAI, some common themes (also in part echoing the practitioners’ view of the first study) emerged: Experts missed clear guidance and standardization on how and when to use XAI. They employed a variety of XAI methods to address specific problems, sometimes successfully (especially as a communication tool between experts) and sometimes unsuccessfully (when data relationships were unknown or too complex). Additionally, XAI methods face several challenges:

- Difficulty in implementation or use.
- Growing complexity of AI systems, such as large language models (LLMs).
- Need to adapt XAI methods to different data types, use cases and AI models, including multi-modal ones.
- Difficulty (or impossibility) in objectively assessing the quality of explanations.

These challenges overall led the experts to believe that XAI is unlikely to provide comprehensive solutions to AI certification in the near future. While some hope was levied towards new XAI-approaches, the most promising methods mentioned regarding AI certification (apart from just spotting ML model biases) were formal verification methods that provide (statistical) guarantees for AI properties and AI approaches incorporating transparent decision-making by design, such as neuro-symbolic approaches.

4 Summary

AI certification and Explainable AI are crucial for ensuring the reliability, safety, and ethical compliance of AI systems. The European Union’s AI Act aims to regulate AI systems by emphasizing transparency, accountability, and risk management, though technical details are still needed. Industry practitioners and researchers can contribute valuable insights for practical frameworks and best practices. Two German studies highlight the need for detailed guidance, robust standards, and trust-building in AI certification. While XAI methods are sometimes mentioned as a possible solution to AI certification, they cannot fully guarantee AI safety but can be valuable tools for identifying biases. As such, they can be used as additional assets in development and certification processes.

References

- 1 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: <http://data.europa.eu/eli/reg/2024/1689/oj>.
- 2 Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development, 2023. [arXiv:2307.11525](https://arxiv.org/abs/2307.11525), doi:10.48550/arXiv.2307.11525.

- 3 Benjamin Fresz, Vincent Philipp Göbels, Safa Omri, Danilo Brajovic, Andreas Aichele, Janika Kutz, Jens Neuhüttler, and Marco F. Huber. The Contribution of XAI for the Safe Development and Certification of AI: An Expert-Based Analysis, 2024. [arXiv:2408.02379](#), doi:10.48550/arXiv.2408.02379.
- 4 Yueqi Li and Sanjay Goel. Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Information Systems Frontiers*, 2024. doi:10.1007/s10796-024-10508-8.
- 5 Jakob Mökander. Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society*, 2(3), 2023. doi:10.1007/s44206-023-00074-y.
- 6 Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B. Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss, and Stefan Wrobel. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog, 2023. [arXiv:2307.03681](#), doi:10.48550/arXiv.2307.03681.
- 7 Joyce Zhou and Thorsten Joachims. How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 12–21, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3593013.3593972.

AI Certification: An Accreditation Perspective

Susanne Kuch¹ ✉ 🏠

Deutsche Akkreditierungsstelle (DAkkS), Stabsbereich Akkreditierungsgovernance, Forschung und Innovation, Berlin, Germany

Raoul Kirmes ✉ 🏠

Deutsche Akkreditierungsstelle (DAkkS), Stabsbereich Akkreditierungsgovernance, Forschung und Innovation, Berlin, Germany

Abstract

AI regulations worldwide set new requirements for AI systems, leading to thriving efforts to develop testing tools, metrics and procedures to prove their fulfillment. While such tools are still under research and development, this paper argues that the procedures to perform conformity assessment, especially certification, are largely in place. It provides an overview of how AI product certifications work based on international standards (ISO/IEC 17000 series) and what elements are missing from an accreditation perspective. The goal of this paper is to establish a common understanding of how conformity assessment in general and certification in particular work regarding AI systems.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; General and reference → Computing standards, RFCs and guidelines

Keywords and phrases certification, conformity assessment, market entry, accreditation, artificial intelligence, standard

Digital Object Identifier 10.4230/OASICS.SAIA.2024.14

Category Practitioner Track

Acknowledgements Special thanks are due to Mattis Jacobs for his support with this paper, and to Dominic Deuber and Svenja Reisinger for their feedback.

1 Introduction

Artificial intelligence (AI) applications are increasingly relevant in every business sector. Some AI systems pose a risk to important commodities worthy of protection, such as health, environment or fundamental rights. This rationale drove the European Union's approach in developing Regulation (EU) 2024/1689 (AI Act) [9]. Although the EU AI Act may be the most comprehensive, other governments like the US or China have provisions with different approaches and regulatory depth [11, 10].

Where regulations exist, compliance must be ensured without establishing new trade barriers. Therefore, the World Trade Organization (WTO) demands for processes of conformity assessment and accreditation² [12]. With the accreditation of a conformity assessment body (CAB), it is ensured that its results can be recognized by other WTO member states as equivalent to their own national CAB results. In this way, accreditation supports international trade since it helps companies avoid duplicating costly conformity assessment procedures in other WTO member countries for companies. In order to enable this system, accreditation

¹ Corresponding author

² Accreditation is a third-party attestation with a conformity assessment body (CAB) as object of conformity. The purpose is to ensure a CAB's formal demonstration of its competence, impartiality and consistent operation in performing specific conformity assessment activities. The authority of an accreditation body can be derived from governments. In Europe, accreditation bodies are government-authorized bodies according to Regulation (EC) No 765/2008.



© Susanne Kuch and Raoul Kirmes;
licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görges, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 14; pp. 14:1–14:7

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

bodies are tasked with assessing the competence, impartiality, and independence of conformity assessment bodies in the sense that these CABs perform their work reliably and in a comparable as well as reproducible manner. Thus, all the procedures and methods applied and used by CABs to perform conformity assessment need to be objective, reproducible and comparable. Therefore, one purpose of this paper is to outline how the established system of conformity assessment works in order to ensure comparability and reproducibility and how this can be applied in the context of AI systems – in an international context as well as with reference to the AI Act. The other is to clarify the distinction of conformity assessment and AI system “testing” and “validation”.

This paper introduces the methodology of the international conformity assessment system in section 2 which is based on the internationally accepted standards series of ISO/IEC 17000, the so-called ISO CASCO toolbox. It will be used to subsequently focus in section 3 on the conformity assessment activity of certification. Here, it will be outlined how certifications with focus on AI systems (mainly seen as products) should be developed in general based on the internationally existent system. It will also briefly explain the distinction between conformity assessment and AI system “testing” and “validation” with respect to the development cycle. This will be followed by the specifications of conformity assessment for the European context in accordance with the EU AI Act in section 4. The paper concludes by outlining some paradigmatic gaps at the technical level of AI systems and by directing to areas where further scientific and standardization work is needed from an accreditation point of view.

2 The methodology of conformity assessment

According to ISO/IEC 17000, conformity assessment means the demonstration that specified requirements are fulfilled.³ Conformity assessment must account for all different kinds of *objects of conformity assessment* such as products, processes, organizations, persons, services, data, systems, materials, designs. Hence, a variety of different types of conformity assessment activities exist. These are mainly testing (including calibration and proficiency testing), inspection, validation/verification and certification (of management systems; persons; products, processes and services). While those activities are all different, they follow the same functional approach determined in ISO/IEC 17000, Annex A. In order to ensure the comparability and reproducibility of the results, each activity and the needed procedure to fulfil the functional approach are defined in the respective conformity assessment standard (ISO/IEC 17000 series), with every one of it internationally agreed upon to support the WTO mutual recognition approach of statements of conformity.

Conformity assessment can be performed by three different entities: the organization providing a product or service (first-party), organizations with user interest (second-party) or independent accredited conformity assessment bodies (third-party). The accredited third-party conformity assessment body is the only one that is allowed to conduct certification as special type of conformity assessment activity according to ISO/IEC 17000 [2].

Third-party certification is therefore an important independent evaluation method to demonstrate certain levels of quality, transparency and trust of a product or service, management system or a competent person to other market players and authorities.

³ Those specified requirements are very often determined in standards and less commonly in technical specifications (in terms of regulation (EU) 1025/2012). Seldom are they directly determined within national regulations.

3 General certification scenarios for AI systems

For AI systems, the following certifications⁴ are particularly relevant: management system certifications (ISO/IEC 17021-1) [3] and certifications for products, processes and services (ISO/IEC 17065) [4].

3.1 Management system certification

Organizations may certify their artificial intelligence management system (AIMS) to show other market participants their competency in accordance with ISO/IEC 42001 [6].⁵ To obtain an AIMS certificate, the organization needs an accredited conformity assessment body (CAB) according to ISO/IEC 17021-1, which offers ISO/IEC 42001 certifications based on ISO/IEC 42006⁶. With such an AIMS certification an organization demonstrates that it is competent to set-up and effectively run a management system for AI systems. This kind of certification can serve as useful evidence for tender procedures or business-to-business relationships in international trade.

The objective of an AIMS is to support an organization in ensuring that the AI systems that are in the scope of the AIMS are developed or deployed as intended and with the respective applicable (quality) requirements (often specified in specific AI system standards). An AIMS is especially important and crucial if requirements for the AI technology itself are not (widely) agreed upon (and standardized) due to a lack of scientific basis (e.g. regarding benchmarks). This becomes especially evident in case of complex AI systems (black-box models). The reason is that they have a limited predictive reliability. Thus, an AI management systems enables the monitoring of the environment in which a complex (black-box model) is deployed and can restrict the impacts of such a model's limited predictive reliability. An AIMS achieves this by monitoring the intended operating conditions and managing, for example, to some degree the processes for data input and generated output. By this, an AIMS can to a certain extent indirectly fill the gap of the non-existent widely scientifically agreed methods to assess some technical requirements for a complex AI model.

3.2 Product certification

Product certifications are performed by accredited third-party CABs only according to ISO/IEC 17065 and follow specific schemes according to ISO/IEC 17067 [5]. ISO/IEC 17065 allows combining different conformity assessment activities (CAA) to evaluate specific AI systems or components thereof (e.g. data, AI model, software). Those *objects of conformity assessment* are assessed against specified requirements listed in the applied certification scheme.

The CAB has to decide on a case-by-case basis which conformity assessment activities (CAA) are applicable to perform its client's assignment. The following CAA are expected to be used during a certification process according to ISO/IEC 17065: *inspection* and *audit*, in some cases laboratory *testing* (with defined or standardized metrics) of software and hardware components or *validation* of statements (e.g. transparency, reliability, level of explainability). Through these conformity assessment activities, a third-party CAB will assess whether or not the set requirements of the specific scheme that was applied to the respective client are met.

⁴ Certifications of persons according to ISO/IEC 17024 can be important but are not considered here. The focus is on the needed conformity assessment activities for market access that determines the AI system as a "product".

⁵ This applies as well for providers of no or low-risk AI systems in Europe.

⁶ ISO/IEC 42006 is currently under development and is likely to be published by the beginning of 2025.

14:4 AI Certification: An Accreditation Perspective

■ **Table 1** Overview of AI systems and its components as objects of conformity mapped towards the applicable conformity assessment activities.

Object of conformity assessment (AI system and/or components)	Conformity assessment activity/activities
Organization (of the AI provider or user)	<i>Audit</i> according to ISO/IEC 17021-1 (for AIMS: ISO/IEC 42006)
Dataset	<i>Inspection</i> according to ISO/IEC 17020 <i>Testing</i> according to ISO/IEC 17025
AI model(s)	<i>Inspection</i> according to ISO/IEC 17020 <i>Validation</i> according to ISO/IEC 17029 <i>Testing</i> according to ISO/IEC 17025
Software (for user interaction)	<i>Inspection</i> according to ISO/IEC 17020 <i>Testing</i> according to ISO/IEC 17025
Hardware	<i>Inspection</i> according to ISO/IEC 17020 <i>Testing</i> according to ISO/IEC 17025
AI system (incl. all components)	<i>Certification</i> based on ISO/IEC 17065, including surveillance and monitoring activities

Table 1 shows the AI system and its components as objects of conformity assessment, mapped to the respective conformity assessment activity which can either be used for a certification scheme or as a standalone conformity assessment activity.

3.3 Identifying gaps within conformity assessment

While non-complex (white-box) AI systems (e.g. knowledge-based systems) can be assessed and certified with the various conformity assessment standards mentioned in Table 1 with a certain level of reliability and certainty, complex (black-box model) AI systems (e.g. deep neural networks (DNN)) have to be assessed indirectly by using CAAs. The focus of indirect assessments is on management systems as well as monitoring activities through embedded inspection that form crucial parts of a certification scheme according to ISO/IEC 17067. Due to their abstract design, certification procedures according to ISO/IEC 17065 are capable of handling high levels of complexity⁷ which is why no additional certification standard on this level of the ISO/IEC 17000 series is required for AI systems.

What is necessary instead is the development of specific schemes in accordance with ISO/IEC 17067 and the monitoring tools for complex (black-box) AI systems. Furthermore, to support the conformity assessment activities, there is the need to develop and publish the specific technical standards on specific methods that specify testing or inspection procedures for AI systems through academia and standardization organizations. An example for such missing standards would be component testing, e.g. for datasets (bias) or AI models (robustness or security features). From an accreditation point of view, it is important to have objective measurands that determine for example if there is bias in the data (or what threshold of bias is acceptable) in order to assess the dataset in a comparable manner or what threshold defines an AI model as “robust” and how it is determined. Together with a measurand for data bias or for robustness, the specific testing procedures for bias or robustness need to be standardized to ensure comparability and reproducibility.

⁷ This is already the case for many different products that mix e.g. software with hardware components in highly regulated areas (e.g. for electronic products in medical area or for construction products, etc.).

In general, accreditation sees a subtle yet recognizable distinction between existent development specific testing and validation methods and procedures as applied by the providers and the methods and procedures used for conformity assessment. The latter need to be internationally aligned through standardization by recognized standardization organizations in order to fulfil the WTO mutual recognition requirement. Also the methods and procedures need to be validated and, if applicable, calibrated such that the tools are applicable to many different objects of conformity (AI systems or the specifically tested components) in order to ensure comparability and reproducibility of the results in an objective manner. Such specifications can only be discussed and agreed upon in standardization and need to be scientifically proven as objective measurand. It may also be necessary to determine whether synthetic datasets should be developed and provided for testing procedures according to ISO/IEC 17025 to serve as objective test or reference datasets. Therefore, high priority must be given to the ongoing work in standardization and in academia to define and develop the needed measurands as well as evaluation methods for AI systems and the components.

4 European specifications and needs

With the AI Act in force, specific requirements need to be fulfilled by European AI operators.

Operators of no- or low-risk AI systems (Art. 3, No. 8, AI Act), only need to fulfil transparency requirements. In contrast, operators of high-risk AI systems must meet all requirements in Chapter III, and demonstrate it by using pre-defined conformity assessment procedures. According to Article 43 AI Act, there are two options: *self-declaration* (first-party) or conformity assessment by a notified body (third-party) (as *product certification*). Yet, where the AI system is part of another product (Annex I, especially Section A), the relevant conformity assessment under those legal acts is required.

All of these options align with the New Legislative Framework (NLF)⁸ and reflect procedures outlined in different modules of decision 768/2008/EC. Module A of this decision applies to the self-declaration and module H1⁹ to a notified body. For AI systems listed in Annex I, the modules of these legal acts apply with the preferred conformity assessment standard (of the 17000 series) to be used according to EA-2/17 [8]. Often, this is also ISO/IEC 17065.

Looking at the second option outlined in Article 43 of the AI Act, notified bodies need to perform conformity assessment (product certification according to ISO/IEC 17065). Thus, also in the European context, the general approach of product certification as outlined in section 3.2 remains applicable. However, the legal consequences differ. In the EU, only a positive assessment by a notified body for a high-risk AI system (here as result of a product certification), where required, permits affixing the CE-marking. Furthermore, the certification process is outlined in Annex VII AI Act. It includes the assessment of the quality management system (QMS) in accordance with Art. 17 of the AI Act via the conformity assessment activity *audit* following ISO/IEC 17021-1, and the technical documentation in accordance of Art. 11 of the AI Act regarding the high-risk AI system via an *inspection* in order to evaluate if all requirements of Chapter III are comprised by the technical documentation.

⁸ The NLF was adopted in 2008 in order to establish a common legal framework for placing goods on the internal market and ensure a high quality of those products placed on the market. The NLF consists of regulation (EC) 765/2008, decision 768/2008/EC and regulation (EU) 2019/1020.

⁹ The European Cooperation for Accreditation (EA) declares in its EA Accreditation for Notification (AfN) Project report from July 2024 [7] that the preferred standard for regulation (EU) 2024/1689 (AI Act) is ISO/IEC 17065.

Accordingly, it applies for the European context as well that specific certification schemes according to ISO/IEC 17067 are developed to ensure compliance with the AI Act based on the already existent CAA (ISO/IEC 17000 series). Since many other harmonized regulations may interfere with the AI Act, it may be feasible to develop this within the different economic sectors in order to take the specific sectoral regulations as well as business practices and applications into account. Thus, for a notified body to attest conformity for an operator, these scheme developments are now crucial. For potential EU-specific objective measurands (e.g. for high quality data), it is necessary that first the standardization experts define the technical requirements for the respective objective of conformity assessment (e.g. data), and subsequently determine which kind of conformity assessment activities may need a specification (e.g. specific testing measurand according to ISO/IEC 17025 for “high quality data” or the development of a procedure to provide “reference data” according to ISO 17034 [1]).

5 Conclusion

In conclusion, we can summarize that the general framework for conducting conformity assessment and especially certification is provided by the ISO/IEC 17000 series and it is generally sufficient for complex systems. This paper demonstrated that this also applies for complex (black-box) and non-complex (white-box) AI systems, since both can be assessed based on these conformity assessment activities. Hence, no additional standards specifically targeting the level of the conformity assessment activities (ISO/IEC 17000 series) are required. However, this paper also made clear what is missing. In particular, there is a need to develop sector and technology-specific certification schemes according to ISO/IEC 17067 – both internationally and at the European level. Furthermore, there is still a gap in standardization regarding objective scientifically proven measurands for certain components (e.g. bias in datasets; security features; robustness) and their calibration as well as the applicable reproducible testing methods. Here, there might be even the need to develop European specific standard testing methods. However, this can only be evaluated once the applicable AI system standards to meet the requirements of the AI Act are developed. From an accreditation perspective, there might also be the need to develop objective synthetic test and reference datasets to ensure a high quality of the testing procedures in laboratories and to reliably assess AI model validation procedures. With regard to complex (black-box) AI systems, there is also a lack of scientific basis that determines whether conformity is presumed or not. Thus, more technical work needs to be done here.

On the other hand, it became evident that in all AI system contexts, AIMS certifications play a vital role for maintaining trust in AI systems and can be considered as even more important than other management systems being used more conventionally as quality assurance measure. The reason is that an AI management system allows to control the context of deployment and to some degree the processes of data input and generated output of an AI system. With regard to complex (black-box) AI systems and the limited predictability of their behaviour, AIMS certifications are thus the only trust anchor available. Consequently, the importance and trustworthiness of the system of accreditation and conformity assessment in general, and in particular of those accredited CABs for these AIMS certifications, should not be underestimated.

References

- 1 DIN Media. DIN EN ISO 17034:2017-04 Allgemeine Anforderungen an die Kompetenz von Referenzmaterialherstellern (ISO 17034:2016); Deutsche und Englische Fassung EN ISO 17034:2016.
- 2 DIN Media. DIN EN ISO/IEC 17000:2020-09 - Konformitätsbewertung - Begriffe und allgemeine Grundlagen (ISO/IEC 17000:2020); Dreisprachige Fassung EN ISO/IEC 17000:2020 - ISO/IEC: ISO/IEC 17000:2020.
- 3 DIN Media. DIN EN ISO/IEC 17021-1:2015-11 - Konformitätsbewertung - Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren - Teil 1: Anforderungen (ISO/IEC 17021-1:2015); Deutsche und Englische Fassung EN ISO/IEC 17021-1:2015.
- 4 DIN Media. DIN EN ISO/IEC 17065:2013-01 - Konformitätsbewertung - Anforderungen an Stellen, die Produkte, Prozesse und Dienstleistungen zertifizieren (ISO/IEC 17065:2012); Deutsche und Englische Fassung EN ISO/IEC 17065:2012.
- 5 DIN Media. DIN EN ISO/IEC 17067:2013-12 - Konformitätsbewertung - Grundlagen der Produktzertifizierung und Leitlinien für Produktzertifizierungsprogramme (ISO/IEC 17067:2013); Deutsche und Englische Fassung EN ISO/IEC 17067:2013.
- 6 DIN Media. ISO/IEC 42001:2023-12 - Informationstechnik – Künstliche Intelligenz – Managementsystem.
- 7 European Accreditation. EA Accreditation for Notification (AfN) Project. Report. Last accessed July,30,2024. URL: <https://european-accreditation.org/wp-content/uploads/2023/04/AFN-PROJECT-2024.pdf>.
- 8 European Accreditation. EA Document on Accreditation for Notification Purposes. Last accessed July 30,2024. URL: <https://european-accreditation.org/publications/ea-2-17-m/>.
- 9 European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, 2016/797/EU and 2020/1828/EU (Artificial Intelligence Act). URL: <http://data.europa.eu/eli/reg/2024/1689/oj>.
- 10 The State Council of the People's Republic of China. China moves to support generative AI, regulate applications. Last accessed 2024-07-30. URL: https://english.www.gov.cn/news/202307/13/content_WS64aff5b3c6d0868f4e8ddc01.html.
- 11 White House of the United States of America. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. Last accessed 2024-07-30. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/?utm_source=link.
- 12 WTO. Agreement on Technical Barriers to Trade. Last accessed 2024-07-26. URL: https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm.

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.2024.15

Category Academic Track

Related Version *Previous Version:* <https://ieeexplore.ieee.org/document/10675927>

Previous Version: <https://digitalcollection.zhaw.ch/server/api/core/bitstreams/488591b2-c384-44d1-b412-18c9b47bd47d/content>

Previous Version: <https://ieeexplore.ieee.org/document/10675803>

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].



© Carmen Frischknecht-Gruber, Philipp Denzel, Monika Reif, Yann Billeter, Stefan Brunner, Oliver Forster, Frank-Peter Schilling, Joanna Weng, and Ricardo Chavarriaga; licensed under Creative Commons License CC-BY 4.0

Symposium on Scaling AI Assessments (SAIA 2024).

Editors: Rebekka Görg, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 15; pp. 15:1–15:18

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

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 [17]. 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. [32], 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 [37] 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 [42]. There is also a substantial risk of environmental damage, privacy breaches, and cybersecurity vulnerabilities. Certain AI models, especially those based on deep neural networks, are known to be vulnerable to adversarial attacks, where subtle modifications to input data can cause significant errors in system behaviour [24]. 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 [31, 28, 16, 47, 40], 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. Noteworthy, 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 requirements and project 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 [53]. Current frameworks often address individual risks in isolation, but aggregating these risks in a way that provides a holistic assessment is complex [5]. Practical challenges in responsible AI implementation, such as balancing transparency, fairness, and robustness, highlight the need for integrated frameworks that address diverse AI risks holistically [8]. Explainability, in particular, plays a critical role in risk aggregation, as it enables stakeholders to interpret AI decision-making processes, fostering trust and accountability [52]. 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 [53]. 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. Explainable AI techniques play a crucial role in fulfilling these transparency requirements, as they allow users to understand and interpret how AI systems reach their decisions. This transparency fosters trust and enables

users to make informed choices, even in lower-risk applications where direct oversight may be minimal [20, 52]. 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 [13].

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 [17]. 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 [59].

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. Several ISO/IEC standards [31] are currently being developed to address AI-related aspects, including terminology, performance metrics, data quality, ethics, and human-AI interaction. Some of these standards have already been released, 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 [26]. The IEEE P7000 series [28] addresses the ethical implications of AI technologies. The European standard development organizations (CEN/CENELEC) have been tasked to develop the standards that will be used to assess conformity with the EU AI Act. As part of this process, they will identify and adopt international standards already developed or under development [10]. 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 [35]. The NIST framework [40] 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 [41]. Moreover, DIN/DKE offers comprehensive standardisation recommendations across all AI domains, facilitating a unified language, principles for development and utilisation, and certification [15]. In the field of aviation, the European Union Aviation Safety Agency (EASA) has introduced comprehensive guidelines for the safe utilisation of ML systems [58]. 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 [47]. 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 [4, 3, 7, 56, 48, 39]. 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 [3] 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. [6] and others [36] have advanced the idea of MLOps as an 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

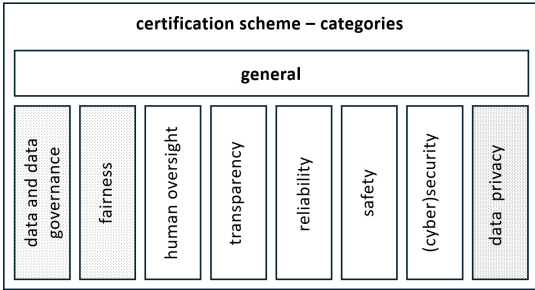
While in some aspects of the verification of AI trustworthiness it is necessary to rely on qualitative results, in particular for model explainability or robustness, automated evaluation workflows mostly involve algorithmic methods with quantifiable output. Therefore, it is crucial to integrate assessment 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 include Microsoft's InterpretML [38], Seldon's Alibi toolbox [54], IBM's AIX360 toolbox [61], Sicara's tf-explain [55], or PyTorch's captum API [9]. Additionally, Quantus [25] 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 [57], Seldon's Alibi-Detect [63], IBM's ART [60] and UQ360 [62] toolboxes. In particular, there are many more toolboxes which implement specific tests for adversarial robustness, such as RobustnessGym [23], CleverHans [46], or Foolbox [49].

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, their 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

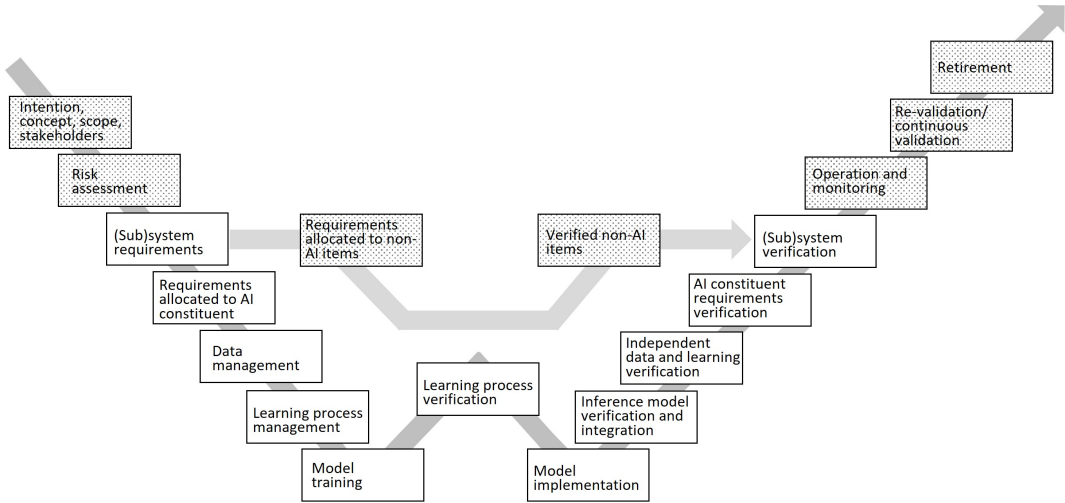
The developed certification scheme for AIS encompasses several principal key aspects of trustworthiness, such as human oversight, transparency, and robustness, which we have defined more granularly by including safety, security, and reliability [14], as illustrated in Figure 1. While we considered the trustworthiness dimensions from the EU, as described in

ALTAI [43], we opted for this more detailed breakdown to enable more precise tracking of objectives. 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 current version of 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 by assessing each AIS as a minimal, limited, or high-risk level application and tailoring the evaluation rigour accordingly. The applicant seeking certification performs and provides the initial risk assessment for their AIS, which then the certifier verifies to ensure alignment with the definitions and criteria set forth in the Act. In addition, risk assessments are based on standards or best practices within different industries and address any additional risks if applicable. The certification process commences with the concept of the system (including the role of the AI part within the overall system), including understanding its role within the larger operational context and determining its 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. In collaboration with the certification body CertX [11], we conducted a targeted analysis of 38 key documents to establish a robust foundation for the certification. These documents, selected in light of the fact that many regulations are still forthcoming and that numerous standards remain under development and are not yet active, were chosen based on their relevance to existing regulations, technical standards, and guidance materials essential for ensuring trustworthiness of AIS. The selection encompassed recognised standard bodies and authoritative guidelines, such as those from the ISO/IEC Joint Technical Committee on Artificial Intelligence (ISO/IEC JTC 1/SC 42) [31], the IEEE Autonomous and Intelligent Systems (AIS) Standards [27] and EASA [58], as well as EU legislative requirements and the Artificial Intelligence Standardization Roadmap developed by DIN and DKE [15], among other sources. 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.

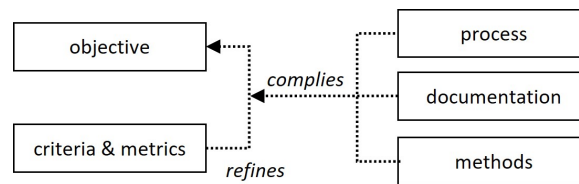


Figure 3 The interrelationship between objectives, criteria and metrics, and compliance methods is illustrated in the diagram. The left side depicts the objectives and their refinement through the application of criteria and metrics, while the right side shows the processes, documentation and methods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

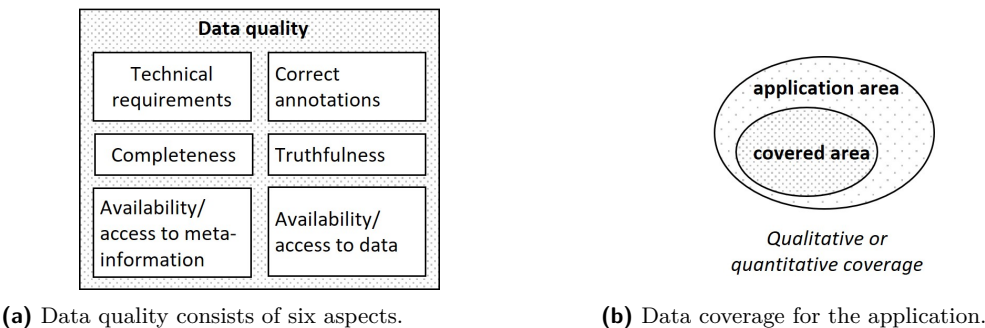
- **MOC:** Provide documentation of methods to provide explanations about the AI/ML item. The type and scope of the provided explanations should be chosen in terms of proportionality, considering the stakeholders.
- **MOC:** Specify the rules that apply to the current decision (e.g., for decision trees, 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 [29, 30, 1, 19]. 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. Guidelines like the NIST AI Risk Management Framework outline methods to minimise bias and ensure fairness [41]. Transparency ensures data is interpretable and preprocessing steps are clear, enabling verification by stakeholders.



■ **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 [44]. 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 [51]. 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 [12]. 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.

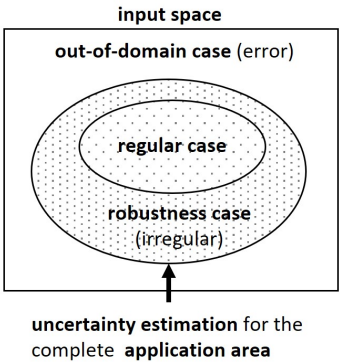
Society: Trust and under-standing by clearly communi-cating the strengths and limitations.	Developers: Clear insights into the internal workings, and limitations.	Users: Transparent explanation of the decisions and results.	Authority: Assurance of regulatory com-pliance and operational transparency by thorough documentation.
	Providers: Monitoring capability by information on internal operations and performance	Auditors & Evaluators: Audit/evaluation capability	

■ **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 [50]. 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 [30, 19]. 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.

Performance metrics			
Regression	likelihood ratio	Confusion matrix	Completeness score
• (Mean) squared error	• True/False-negative rate	• Hinge loss	
• (Mean) absolute error	• True/False-positive rate		Ranking
Classification	• Precision-recall curve	Computer vision	• Mean reciprocal rank
• (Balanced) accuracy	• Receiver operation characteristics (ROC)	• Peak-signal-to-noise ratio (PSNR)	• Discounted cumulative gain
• Micro/macro average	• Lift	• Structural similarity	Natural Language Processing
• (Balanced) F1-score	• Matthew's correlation coefficient	• (Mean) intersection over union (mIOU)	• Perplexity score
• Prevalence	• Area under curve (AUC)	Clustering	• BLEU score
• Precision	• Cohen's Kappa	• Silhouette value	
• False discovery/omission rate		• Adjusted mutual information score	
• Positive/negative			

■ **Figure 7** Non-exhaustive list of performance metrics used for regression, classification, computer vision, clustering, ranking, and natural language processing.

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.

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.

3.2.5 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 [30]. 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 [18]. 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 [24].

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, including adherence to the EU AI Act, standards from ISO/IEC and IEEE [30, 17, 26, 44], among others. To achieve this, we evaluated several requirements management tools and ultimately selected Jira as the tool for organising, documenting, and tracing 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 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.

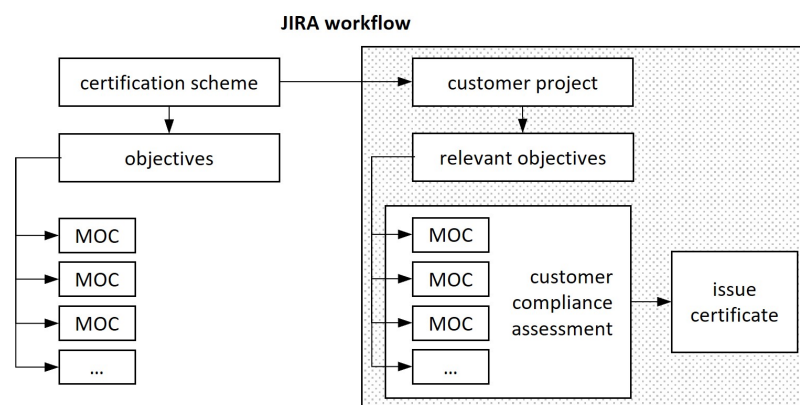
Using Jira for requirement management ensures each objective and MOC is meticulously organised, facilitating clear communication and comprehensive oversight. Its ability to maintain detailed records and provide real-time updates is crucial for this task. Real-time collaboration and review capabilities 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.

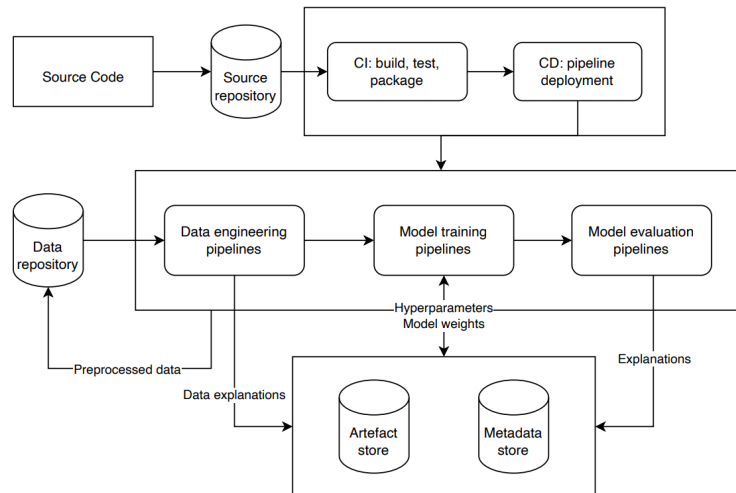
4.2 Machine Learning Operations Infrastructure

As argued in section 2.2, MLOps serves as an 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 [6]. MLOps extends DevOps practices to manage the complexities of bringing AIS into production, ensuring that they continuously meet trustworthiness standards [33].

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.



■ **Figure 9** Overview of the MLOps system architecture.

In our pipeline, models are developed, trained and versioned using Git (through GitHub [21]) and MLflow [34], 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 [22] and Apache Airflow [2] are used for workflow scheduling and monitoring and facilitate automated testing and deployment processes in CI/CD pipelines. Data is versioned using Oxen [45]. 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 3.2.5. 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 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 [14]. 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 [64].

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 IEEE 7001-2021 - IEEE Standard for Transparency of Autonomous Systems. Technical report, Institute of Electrical and Electronics Engineers, 2021. URL: <https://standards.ieee.org/standard/7001-2021.html>.
- 2 Apache Software Foundation. Airflow. URL: <https://airflow.apache.org/>.
- 3 Atlassian. Jira, 2002. URL: <https://www.atlassian.com/software/jira>.
- 4 Atlassian. Confluence, 2004. URL: <https://www.atlassian.com/software/confluence>.
- 5 Richard Benjamins, Alberto Barbado, and Daniel Sierra. Responsible AI by design in practice. *arXiv preprint arXiv:1909.12838*, 2019.
- 6 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.
- 7 Jace Browning and Robert Adams. Doorstop: Text-based requirements management using version control, 2014. doi:10.4236/jsea.2014.73020.
- 8 Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

- 9 Captum. Model interpretability for PyTorch, 2023. URL: <https://captum.ai/>.
- 10 CEN-CENELEC. Artificial Intelligence. URL: <https://www.cenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/>.
- 11 CertX. CertX: First Swiss Functional Safety and Cyber Security Certification Body. URL: <https://certx.com/>.
- 12 Chun Sik Chan, Huanqi Kong, and Guanqing Liang. A comparative study of faithfulness metrics for model interpretability methods. *arXiv preprint arXiv:2204.05514*, 2022. doi:10.48550/arXiv.2204.05514.
- 13 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/>.
- 14 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.
- 15 DIN, DKE. Artificial intelligence standardization roadmap, 2023. URL: <https://www.dke.de/en/areas-of-work/core-safety/standardization-roadmap-ai>.
- 16 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>.
- 17 European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 9 october 2024 on harmonized rules for artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- 18 International Organization for Standardization. ISO/IEC 27001:2013 information technology – security techniques – information security management systems – requirements. Technical report, ISO, 2013.
- 19 International Organization for Standardization. ISO/IEC 24029-1:2021 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview. Technical report, International Organization for Standardization, 2021. URL: <https://www.iso.org/standard/77609.html>.
- 20 Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. doi:10.1109/DSAA.2018.00018.
- 21 GitHub, Inc. GitHub . URL: <https://github.com/>.
- 22 GitHub, Inc. GitHub Actions. URL: <https://github.com/features/actions>.
- 23 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>, doi:10.18653/V1/2021.NAACL-DEMOS.6.
- 24 Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. 2014. doi:10.48550/arXiv.1412.6572.
- 25 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>.
- 26 IEEE. IEEE CertifAIED: the mark of AI ethics, 2022. URL: <https://engagestandards.ieee.org/ieeecertifaiied.html>.

- 27 IEEE Standards Association. IEEE Autonomous and Intelligent Systems Standards. URL: <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/>.
- 28 IEEE Standards Association. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2023. URL: <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>.
- 29 International Organization for Standardization. ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. Technical report, 2015. URL: <https://www.iso.org/standard/35746.html>.
- 30 International Organization for Standardization. ISO/IEC 24028:2020 Information technology — Artificial intelligence (AI) — Overview of trustworthiness in AI. Technical report, 2020. URL: <https://www.iso.org/standard/77608.html>.
- 31 ISO. ISO/IEC JTC 1/SC 42 Artificial Intelligence, 2023. URL: <https://www.iso.org/committee/6794475.html>.
- 32 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.
- 33 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.
- 34 LF Projects, LLC. MLFlow. URL: <https://mlflow.org/>.
- 35 LNE. Certification of processes for AI, 2023. URL: <https://www.lne.fr/en/service/certification/certification-processes-ai>.
- 36 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.
- 37 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021. doi:10.1145/3457607.
- 38 Microsoft. InterpretML. URL: <https://github.com/interpretml/interpret>.
- 39 Microsoft. Azure devops, 2005. URL: <https://azure.microsoft.com/en-us/products/devops/#overview>.
- 40 NIST. NIST Technical AI Standards, 2023. URL: <https://www.nist.gov/artificial-intelligence/technical-ai-standards>.
- 41 NIST. AI Risk Management Framework (AI RMF) Knowledge Base, 2024. URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.
- 42 Office of the United Nations High Commissioner for Human Rights (OHCHR). A taxonomy of AI and human rights harms. Technical report, United Nations Human Rights Office of the High Commissioner, 2023. Accessed: 2024-11-06. URL: <https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf>.
- 43 High-Level Expert Group on Artificial Intelligence. Assessment list for trustworthy artificial intelligence (ALTAI), 2020. URL: <https://altai.insight-centre.org>.
- 44 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>.
- 45 Oxen.ai. oxen. URL: <https://www.oxen.ai/>.
- 46 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.

- 47 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>.
- 48 Rational Software. IBM doors, 2018. URL: <https://www.ibm.com/docs/en/engineering-lifecycle-management-suite/doors>.
- 49 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.
- 50 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/.
- 51 Wojciech Samek et al. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. doi:10.1007/978-3-030-28954-6.
- 52 Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.
- 53 Anna Schmitz, Michael Mock, Rebekka Görges, Armin B Cremers, and Maximilian Poretschkin. A global scale comparison of risk aggregation in AI assessment frameworks. *AI and Ethics*, pages 1–26, 2024.
- 54 Seldon. Alibi Explain. URL: <https://github.com/SeldonIO/alibi>.
- 55 sicara. TF-Explain: Interpretability Methods for tf.keras Models with TensorFlow 2.x. URL: <https://github.com/sicara/tf-explain>.
- 56 Siemens. Polarion, 2004. URL: <https://polarion.plm.automation.siemens.com/>.
- 57 Ryan Soklaski, Justin Goodwin, Olivia Brown, Michael Yee, and Jason Matterer. Tools and practices for responsible AI engineering. *arXiv preprint arXiv:2201.05647*, 2022. arXiv: 2201.05647.
- 58 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>.
- 59 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/>.
- 60 Trusted-AI LF AI Foundation. Adversarial Robustness Toolbox (ART). URL: <https://github.com/Trusted-AI/adversarial-robustness-toolbox>.
- 61 Trusted-AI LF AI Foundation. AI Explainability 360 (AIX360). URL: <https://github.com/Trusted-AI/AIX360>.
- 62 Trusted-AI LF AI Foundation. AI Uncertainty Quantification 360 (UQ360). URL: <https://github.com/Trusted-AI/UQ360>.
- 63 Arnaud Van Looveren et al. Alibi detect: Algorithms for outlier, adversarial and drift detection, 2019. URL: <https://github.com/SeldonIO/alibi-detect>.
- 64 Larysa Visengeriyeva, Anja Kammer, Isabel Bär, Alexander Kniesz, and Michael Plöd. MLOps Principles, 2020. URL: <https://ml-ops.org/content/mlops-principles>.