Trustworthy AI

WHITE PAPER

Evaluation methods and frameworks

Sebastian Scher*, Simone Kopeinik*, Christof Wolf-Brenner*, Kerstin Waxnegger *,
Tomislav Nad** and Dominik Kowald*

October 2024

* Know Center Research GmbH
** SGS Digital Trusts Services GmbH

Partners of SGS

KNOW Center

SGS

Contents:

Introduction

Evaluation of artificial intelligence (AI) is an increasingly relevant research topic. While there is currently no universal consensus on what constitutes evaluation of AI, the concept of evaluation itself is defined as the process of assessing the merit and worth of an object [Stufflebeam01]. In line with this definition, we can say that evaluation of AI is the process of assessing the merit and worth of AI.

Recently, there has been an increasing interest in the trustworthiness of AI. To correctly evaluate AI in this regard, it is essential to view an AI application as a system comprised of several components that interact with each other [SJH+20]. The three main components that most AI systems have are algorithms, data and computing infrastructure [ACJ20]. Hence, a complete evaluation of an AI system's trustworthiness should consider all these components, as well as any additional components of the concrete AI system to be assessed. Since components of an AI system interact with each other, it is crucial to not only evaluate each component in isolation but also to assess components jointly with respect to their interaction.

There is no single method for evaluating an AI system (and its components). The trustworthiness of an AI system depends on several dimensions, such as fairness, reliability and safety, and each dimension is typically evaluated by a dedicated evaluation method. To assess trustworthiness in an objective and reproducible way, it is necessary to use an evaluation framework, which is a bundle of evaluation methods for several dimensions of trust, and for all components in an AI system.

An **evaluation framework** covers all components and all stages of the **AI** lifecycle.

Existing frameworks for evaluating the trustworthiness of AI

In recent years, several competing frameworks for evaluating AI trustworthiness were proposed. These frameworks are highly detailed and diverse with respect to which dimensions of trustworthiness, which stages of the AI lifecycle and which components are covered. Moreover, the concrete evaluation methodology and the level of technical detail also differ significantly between frameworks.

Hence, it is not obvious which framework can be suitable in a practical evaluation task. To bring more clarity to this issue, this document discusses a selection of published frameworks.

At the time of writing of this article, some of the most prominent examples for evaluation frameworks are:

Name	Date published	Description	Domain	Reference
Fraunhofer Catalog	2021 (German); 2023 (English);	Audit catalog for assessing risks of AI systems	Domain-independent	[FH23]
capAI	2022	Document for establishing compliance with EU Artificial Intelligence Act (AIA)	Domain-independent	[FHT+22]
ALTAI	2020	Self-assessment list published by European Commission	Domain-independent	[ABB+20]
VDE-SPEC	2022	Specification of a standard for establishing compliance with AIA	Domain-independent	[LHP+22]
examAI	2021	Document for delineation of AI systems as socio-technical systems	Domain-independent, but focus on Working environment	[WB21]
TÜV Austria	2021	Audit catalog for low-risk AI applications	Domain-independent	[WEW+21]
FUTURE-AI	2021	Framework for assessing AI systems in medical imaging	Medicine	[LOG+21]

Each of these frameworks consists of a varyingly detailed methodology and a varying number of technical details. While all frameworks mention numerous relevant dimensions of trustworthiness, not all include sufficient detail to enable an auditor to adequately assess all dimensions.

The Fraunhofer catalog [FH23] is a structured guideline for concretizing abstract quality standards into AI application specific criteria. This guideline can be used for evaluating AI systems with respect to trustworthiness in an audit-like manner.

In terms of methodology, the catalog focuses on the evaluation of specific risks related to AI and its trustworthiness, e.g. the risk of input corruption leading to unreliable AI behavior. These specific risks are assessed and combined into wider and more general risk assessments to obtain an overall assessment of a dimension of trustworthiness, as well as cross-dimensional risk assessments. The AI-developer needs to show that sufficient measures have been taken to reduce the risks to an acceptable level.

> The conformity assessment procedure **capAI [FHT+22]** is a document for ensuring and demonstrating that an **AI** system conforms to the AIA.

It has a focus on demonstrating trustworthiness via legal compliance, ethical soundness and technical robustness, and implements ethics-based auditing principles to several stages of the AI lifecycle. The conformity assessment procedure consists of three separate documents: First, an internal review protocol for producing documentation that is required by the AIA; Second, a summary datasheet, which contains a high-level overview of the characteristics of the AI system with respect to legal requirements; Third, an external scorecard for communicating the AI systems characteristics and its level of trustworthiness to stakeholders.

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) [ABB+20] is a self-assessment list published by the European Commission. It is aimed at encouraging thoughtful reflection to provoke appropriate action and nurture an organizational culture committed to maintaining trustworthy AI systems. The list consists of 135 yes-or-no questions for self-assessing the trustworthiness of an AI system.
To answer these questions, it is recommended to construct a multi-disciplinary team of:

- AI designers
- Data scientists
- Procurement officers
- Front-end staff that will work with the AI system
- Legal/compliance officers
- Management

ALTAI aims at leveraging the expertise of this multi-disciplinary team to obtain an adequate self-assessment of the trustworthiness of an organization's AI system.

VDE-SPEC [LHP+22] is the specification of a standard designed to be compatible with the AIA developed by the German "Verband der Elektrotechnik". This specification is comparable to other German standards for making product qualities transparent, e.g. VDE-SPEC is similar to the Nutri-Score standard which enables consumers to easily assess the nutrition value of an edible product.

To assess the trustworthiness of AI, VDE-SPEC uses a value-criterion-indicator-objective model paired with a question catalog and a rating-aggregation system. The state of an AI system with respect to an aspect of trust is documented and associated with an easily interpretable rating on a scale from A (best) to G (worst). The rating aggregation system the combination of ratings of individual aspects of trust to an overall rating for the entire AI system.

The framework examAI [WB21] has the objective of achieving a complete delineation of AI systems as socio-technical systems. It has a special focus on auditing AI applications in working environments and characterizes AI via the following components: algorithms, data and computing infrastructure. In terms of applications, it defines the key focus topics of talent and human resource management and describes use cases for auditing the trustworthiness of AI in such an application. The framework also contains a technical section which explains how the auditing process interacts with technical aspects of AI.

The TÜV Austria framework [WEW+21] is an audit catalog for low-risk AI applications such as automated processing of non-critical documents. The catalog consists of roughly two hundred requirements and dedicated sections on the basics of standardization, certification and on the technical background. There is also a list of eleven major challenges for the certification of low-risk machine learning applications. The main objectives of the framework are the following: First, to clarify important machine learning principles in simple terms in order to reach a broad audience; Second, to discuss important machine learning-related aspects and challenges in the context of certification; Third, to utilize existing certification procedures to take a first step towards developing a certification procedure for machine learning applications. The details of the framework's methodology are not publicly available.

FUTURE-AI is an evaluation framework specifically designed for the medical domain. It introduces a careful selection of guiding principles drawn from the accumulated experiences, consensus and best practices from five large European projects on AI in health imaging. The framework's name is an acronym derived from the dimensions Fairness, Universality, Traceability, Usability, Robustness and Explainability. In terms of methodology, it contains a set of 55 questions for assessing trustworthiness of a medical AI system. Due to the domain specific nature of FUTURE-AI, it also contains special steps such as clinical conceptualization. The framework is also available as an online version.

Related frameworks

In addition to the aforementioned frameworks, several (proposed) frameworks focus on assessing and mitigating risks of AI systems. In our opinion, these frameworks should also be considered when evaluating the trustworthiness of AI.

For example, the NIST AI Risk Management Framework (NIST AI RMF) [NI24] is a structured framework designed to help organizations manage risks associated with the development and deployment of AI systems, with a particular focus on ensuring trustworthiness. The framework aims to provide a comprehensive approach to identifying, assessing and mitigating risks across the entire AI lifecycle, from design and development to deployment and decommissioning. By focusing on key dimensions of trustworthiness such as safety, security, fairness and transparency, the NIST AI RMF attempts to enable organizations to align AI-related activities with their risk tolerance, regulatory requirements and societal expectations. The framework emphasizes the importance of ongoing monitoring and adaptation to evolving risks, ensuring that AI systems remain trustworthy over time. This framework aims to empower organizations to systematically address both known and emerging risks, including those exacerbated by specific AI technologies like generative AI.

Also exemplarily, the proposed Human Rights, Democracy and the Rule of Law Assurance Framework (HUDERAF) is an algorithmic-neutral, practice- and risk-based approach to assessing and mitigating adverse impacts developed for the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. The proposal was designed in collaboration between the Council of Europe's Committee on AI and The Alan Turing Institute and encompasses a Preliminary Context-Based Risk Analysis (PCRA), a Stakeholder Engagement Process (SEP), a Human Rights, Democracy and the Rule of Law Impact Assessment (HUDERIA) and a Human Rights, Democracy and Rule of Law Assurance Case (HUDERAC). It follows a comprehensive approach to responsible AI governance by considering both the technical aspects of AI systems and the sociotechnical context in which they are developed and applied. [CAHAI21] [LBA+22]

Dimensions of trustworthiness

There are several dimensions of trustworthiness that are addressed in the discussed frameworks. These dimensions are sometimes referred to under different names, or the frameworks use similar terms to refer to different dimensions. To bring more clarity to this issue, this document contains a harmonized list of dimensions based on all material found in the discussed frameworks. The content and relevancy of each dimension is illustrated in the following.

Fairness

A framework for evaluating the trustworthiness of an AI system typically provides the means to determine whether the system under consideration is fair. The AI system is assessed for compliance with ethical and/or legal requirements that delimit the boundaries of morally acceptable behavior within a specific society.

These requirements can be equal treatment of all individuals, equity, between (privileged and disadvantaged) groups and others. From a technical point of view, these requirements commonly translate to restrictions on permitted outputs/decisions and how they are influenced with respect to sensitive attributes. For example, an AI system for predicting whether a convicted felon will be arrested again after their release may not use age or skin complexion as means for making this prediction (see [Rudin19] for a detailed discussion of the deficiency of such systems).

Autonomy and control

Several evaluation frameworks provide the means to assess the risk of an AI system's environment and determine which degree of autonomy is permissible, and whether there are sufficient control mechanisms in place. Such an evaluation is highly beneficial, as AI systems frequently operate without direct human supervision. For instance, they may be considered for operation in high-risk environments such as power plants or medical environments where the application of fully autonomous AI systems may not be appropriate or even prohibited by law.

Furthermore, an AI system may also limit a human's autonomy, for example, by restricting access to sensitive documents or inappropriate content. Depending on the setting, this may or may not have legal implications that need to be evaluated accordingly. The degree to which existing frameworks cover this dimension of trustworthiness is hard to assess since relevant text passages are frequently distributed across the document describing the framework.

Environmental and societal impact

Depending on the application, an AI system may have a significant impact on the environment. An AI system consumes energy and may further require data centers or similar large processing facilities that also consume resources. Hence, several evaluation frameworks for AI systems also cover this aspect of trustworthiness – whether we can trust the AI system to be sustainable and to handle limited resources in a responsible manner [YC20]. This is a facet of trust that is often overlooked, although it is a central issue in many contemporary discussions. Part of these is the ever-increasing amount of computing power [OAI 18] – and correspondingly energy consumption – needed for AI models. However, different types of AI systems have vastly different energy requirements [KDS22], end especially large language models have become scrutinized for their high power consumption [EAA19].

An AI system may also have a wider societal impact that may be assessed by an evaluation framework. For instance, if an AI system may amplify fake news or facilitate totalitarian behaviors [ABB+20].

If such an impact is overlooked during evaluation, an AI system's trustworthiness might be falsely assessed. This explains why several evaluation frameworks also cover the wider societal impact of an AI system.

Transparency

Since the rise of deep learning-based AI systems, the issue of transparency has become increasingly relevant. Due to the "black box" nature of such systems, it is difficult for humans to understand how the AI operates and whether it complies with legal and ethical standards. As an extreme example, one can consider the COMPAS (Correlational Offender Management Profiling for Alternative Sanctions) system applied by the US justice system. COMPAS was a complex proprietary system for assessing recidivism risk based on 137 variables such as typographical data and questionnaire responses [RWC20]. While widely used, some researchers and activists gained increasing suspicions that this untransparent proprietary AI system treats individuals highly unfairly. Finally, a research group could show that the complicated COMPAS AI system was essentially behaving like a simple rule list [Rudin19]:

IF	Age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	Age between 21-23 and 2-3 prior offences	THEN predict arrest
ELSE IF	More than three priors	THEN predict arrest
ELSE	Predict no arrest	

Clearly, this prediction mechanism conflicts with moral standards and the law. If the system had been as transparent as the rule list above, it would likely never have been used in an official justice system.

Hence, an AI system must be transparent, and every responsible assessment of an AI system must evaluate its transparency. Accordingly, it is beneficial if AI evaluation frameworks cover and assess transparency as a dimension of trustworthiness. One of the main prerequisites for an AI system to be transparent is that the decisions that the systems make can be understood or explained. This is the topic of the highly active research field of explainability of AI – or Explainable AI (XAI) [GSC+19].

> This field researches methods to explain the decisions of **AI** systems, and thus to overcome the black-box nature of **AI** systems.

Reliability

Evaluation frameworks typically also cover reliability. In simple terms, an AI system's trustworthiness is directly affected by how reliably it operates. In more detail, this implies certain robustness criteria that need to be assessed, such as robustness to adversarial attacks, resilience to outliers and missing data. In general, reliability is one of the widest dimensions of trustworthiness of AI and is difficult to cover completely without leaving any gaps open. Some well-established procedures for assessing reliability are penetration tests [ABB+20] and quantification via performance measures [FH23]. However, there are many further aspects to consider.

For a wholesome assessment of reliability, it is desirable that a framework has an adequate technical section that equips the evaluator with a sufficient understanding of the technical details. Otherwise, the evaluation of reliability is bound to be too superficial to cover this dimension in its entirety.

Security and safety

An improperly assessed AI system may pose security risks. The attack surface of AI systems is usually large, and an adversary has many avenues for an attack attempt. Clearly, the trustworthiness of an AI system is linked to security – how can one trust a system whose integrity cannot be assured? Likewise, it is difficult to trust an AI system if one cannot guarantee that it will be available in a critical situation, e.g. because of a denial-of-service attack [CKG+06]. Hence, there is ample motivation covering security and related aspects in an AI trustworthiness evaluation framework.

Similarly, safety is also a highly relevant issue. In some scenarios, e.g. a robot-arm operated by AI, an AI system may pose risks to an individual's physical well-being or may cause damage to property. The assessment of safety-related aspects is often driven via worst-case analyses [PJP07]. For example, a safety assessment of the above-mentioned robot arm may revolve around determining the maximum possible harm that this arm could cause during operation, and which measures are in place to keep this maximum small.

Data protection

Data privacy is a fundamental right granted by the EU General Data Protection Regulation and similar regulations in many other countries. Thus, it is natural that many evaluation frameworks assess an AI system's capability of ensuring data privacy and its compliance with such regulations. In an era where "big business" can be made with personal data, it is essential to balance AI's increasing demand for data and personal rights [MM19]. Likewise, this trade-off is crucial when evaluating AI systems. On the one hand, compliance with regulations is a must and any violation should be visible. On the other, one must balance the evaluation so that it is not too harsh that a legal AI system is not permitted the necessary minimum of data to operate properly. Existing frameworks vary considerably in this regard and are not all equally successful at achieving the aforementioned trade-off.

Summary

There is an increasing interest in the evaluation of AI trustworthiness. To properly evaluate the trustworthiness of AI, one needs an entire evaluation framework consisting of many evaluation methods, where each method addresses a specific aspect of trust. We have provided an overview of seven frameworks for evaluating the trustworthiness of AI as well as discussed two related frameworks. The frameworks vary considerably in level of detail, coverage and completeness. According to the best current knowledge, the dimensions of AI trustworthiness can be harmonized to a set of seven dimensions and it is essential that the evaluator using such a framework is equipped with the required technical expertise. Further study is needed to assess the level of detail and completeness of each framework's methodology and technical content. Finally, it must be pointed out that it is a delicate matter to responsibly assess how well existing frameworks cover a dimension and whether the provided technical details are sufficient, too vague or too extensive. It will only be possible to answer this via applying the frameworks to many different AI applications and use-cases.

References

- [AAA+13] Al-Kharabsheh, K. S., AlTurani, I. M., AlTurani, A. M. I., & Zanoon, N. I. (2013). Review on sorting algorithms a comparative study. *International Journal of Computer Science and Security* (IJCSS), 7(3), 120-126
- [ABB+20] Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A. (2020). The assessment list for trustworthy artificial intelligence (ALTAI). European Commission
- [ALC20] van Assen, M., Lee, S. J., & De Cecco, C. N. (2020). Artificial intelligence from A to Z: from neural network to legal framework. European journal of radiology, 129, 109083
- [CAHAI21] Ad Hoc Committee On Artificial Intelligence, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems (HUDERAF) (2021)
- [CKB+06] Carl, G., Kesidis, G., Brooks, R. R., & Rai, S. (2006). Denial-of-service attack-detection techniques. *IEEE* Internet computing, 10(1), 82-89
- [Chollet21] Chollet, F. (2021). Deep learning with Python. Simon and Schuster
- [FFG+20] Fetic, L., Fleischer, T., Grünke, P., Hagendorf, T., Hallensleben, S., Hauer, M., ... & Puntschuh, M. (2020). From Principles to Practice. An interdisciplinary framework to operationalise AI ethics.
- [EAA19] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243
- [FHT+22] Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. Available at SSRN 4064091
- [FH23] Fraunhofer IAIS (2023), Guideline for Designing Trustworthy Artificial Intelligence AI Assessment Catalog, https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/Fraunhofer_IAIS_AI_ASSESSMENT_Catalog_Web.pdf
- [GC22] Gabidolla, M., & Carreira-Perpiñán, M. Á. (2022). Optimal interpretable clustering using oblique decision trees. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 400-410)
- [GSC+19] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science robotics, 4(37)
- [Huber04] Huber, P. J. (2004). Robust statistics (Vol. 523). John Wiley & Sons
- [HCM+22] Huang, A., Chao, Y., de la Mora Velasco, E., Bilgihan, A., & Wei, W. (2022). When artificial intelligence meets the hospitality and tourism industry: an assessment framework to inform theory and management. Journal of Hospitality and Tourism Insights, 5(5), 1080-1100
- [HUD21] ad hoc committee on artificial intelligence (2021), Human Rights, Democracy, and the Rule of Law Assurance Framework (HUDERAF), https://rm.coe.int/cahai-pdg-2021-09-huderaf-executive-summary/1680a416de
- [KDS+22] Kaack, L.H., Donti, P.L., Strubell, E. et al. Aligning artificial intelligence with climate change mitigation. Nat. Clim. Chang. 12, 518–527 (2022)
- [LBA+22] Leslie, D., Burr, C., Aitken, M., Katell, M., Briggs, M., & Rincon, C. (2022). Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. Zenodo. https://doi.org/10.5281/zenodo.5981676
- [LHP+22] Loh, W., Hauschke, A., Puntschuh, M., & Hallensleben, S. (2022). VDE SPEC 90012 V1. 0: VCIO Based Description of Systems for AI Trustworthiness Characterisation. https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf
- [LOG+21] Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., ... & Martí-Bonmatí, L. (2021). FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv preprint arXiv:2109.09658
- [MAC+22] Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Minds and Machines, 32(2), 241-268
- [MM19] Mazurek, G., & Małagocka, K. (2019). Perception of privacy and data protection in the context of the development of artificial intelligence. Journal of Management Analytics, 6(4), 344-364
- [NI23] National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0)
- [OAI18] AI and Compute, OpenAI, 2018, https://openai.com/research/ai-and-compute
- [PJP07] Prajna, S., Jadbabaie, A., & Pappas, G. J. (2007). A framework for worst-case and stochastic safety verification using barrier certificates. *IEEE* Transactions on Automatic Control, 52(8), 1415-1428
- [Rudin19] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215
- [Stufflebeam01] Stufflebeam, D. (2001). Evaluation models. New directions for evaluation, 2001(89), 7-98
- [RWC20] Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 2(1), 1
- [SJH+20] Siebert, J., Joeckel, L., Heidrich, J., Nakamichi, K., Ohashi, K., Namba, I., ... & Aoyama, M. (2020). Towards guidelines for assessing qualities of machine learning systems. In Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9–11, 2020, Proceedings 13 (pp. 17-31). Springer International Publishing
- [WB21] Waltl, B., Becker, N. (2021). KI-Audit in der Arbeitswelt: Ein integratives Framework zum Auditieren und Testen von KI-Systemen. https://gi.de/fileadmin/PR/Testing-AI/ExamAI_Framework_KI-Audit.pdf
- [WEW+21] Winter, P. M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., ... & Nessler, B. (2021). Trusted artificial intelligence: Towards certification of machine learning applications. arXiv preprint arXiv:2103.16910
- [YC20] Yigitcanlar, T., & Cugurullo, F. (2020). The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities. Sustainability, 12(20), 8548
- [YW22] Yang, R., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. Electronic Markets, 32(4), 2053-2077
- [ZTK+19] Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., & Caruana, R. (2019, July). Axiomatic interpretability for multiclass additive models. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 226-234)

When you need to be sure

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

🌐 https://know-center.at/

✉ info@know-center.at