*Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)*
*Edited by* Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio

doi: 10.3850/978-981-18-8071-1_P230-cd

# A Comprehensive Framework for Ensuring the Trustworthiness of AI Systems

Stefan Brunner, Carmen Mei-Ling Frischknecht-Gruber, Monika Reif, Joanna Weng
*Institute of Applied Mathematics and Physics, Zurich University of Applied Sciences, Switzerland.*
*E-mail: stefan.brunner@zhaw.ch, carmen.frischknecht-gruber@zhaw.ch, monika.reif@zhaw.ch, joanna.weng@zhaw.ch*

Legislators and authorities are working to establish a high level of trust in AI applications as they become more prevalent in our daily lives. As AI systems evolve and enter critical domains like healthcare and transportation, trust becomes essential, necessitating consideration of multiple aspects. AI systems must ensure fairness and impartiality in their decision-making processes to align with ethical standards. Autonomy and control are necessary to ensure the system remains aligned with societal values while being efficient and effective. Transparency in AI systems facilitates understanding decision-making processes, while reliability is paramount in diverse conditions, including errors, bias, or malicious attacks. Safety is of utmost importance in critical AI applications to prevent harm and adverse outcomes. This paper proposes a framework that utilizes various approaches to establish qualitative requirements and quantitative metrics for the entire application, employing a risk-based approach. These measures are then utilized to evaluate the AI system. To meet the requirements, various means (such as processes, methods, and documentation) are established at system level and then detailed and supplemented for different dimensions to achieve sufficient trust in the AI system. The results of the measures are evaluated individually and across dimensions to assess the extent to which the AI system meets the trustworthiness requirements.

*Keywords*: Artificial Intelligence, Trustworthiness of AI systems, AI Standards, AI Safety.

## 1. Introduction

As Artificial Intelligence (AI) has become an increasingly important part of our lives, impacting industries and society as a whole, concerns about the safety and reliability of these systems have also increased. In April 2021, the European Union (EU) proposed new legislation on AI, which aims to establish a comprehensive regulatory framework for AI systems (AIS) in the EU (Council of European Union, 2021). One of the key provisions of this legislation is a risk-based approach that divides AIS into four categories based on the potential harm they may cause. Thus, it prohibits certain AI practices considered ”unacceptable” and pose a significant risk to fundamental rights. The EU legislation imposes various requirements on high-risk AIS, encompassing safety, reliability, transparency, human oversight, and accountability. For limited-risk AIS, ensuring transparency is essential to foster trust, accountability, and informed decision-making among stakeholders such as users, developers, and regulators.

## 2. Background

Numerous national and international organizations are involved in initiatives to promote trust in AI. The LNE’s AI certification program establishes impartial and objective criteria for trustworthy AIS, including ethics, safety, transparency, and privacy (LNE, 2023). IEEE is developing a certification program to assess the transparency, accountability, bias, and privacy of AI-related processes (IEEE, 2022). EASA has published a comprehensive guideline for the safe use of Machine Learning (ML) in the aviation sector (Soudain, 2021). This guideline aims to support stakeholders in developing and implementing ML systems with low levels of automation, covering the entire lifecycle from development processes to the use of ML in operations. DIN/DKE provides detailed recommendations for standardization across all AI topics to establish a common language, principles for development and use, and certification (DIN, DKE, 2023). The Fraunhofer Institute has developed a guideline for designing trustworthy AI (Poretschkin et al., 2021) to increase trust in AIS. This AI catalog evaluates the trustworthiness of AIS based on six dimensions: fairness, autonomy and control, transparency, reliability, safety/security, and privacy. In contrast to other contributions, the guideline proposes several technical methods in addition to process-related measures for the evaluation of AIS.

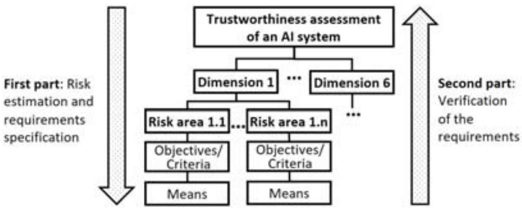



Fig. 1.: Framework by Poretschkin et al. (2021).

The framework considers six dimensions: Fairness, Reliability, Transparency, Autonomy and Control, Safety and Security, and Data Privacy. The AI catalog divides the evaluation process into mainly two parts where the first one consists of the risk estimation and requirements specification of the AIS for a specific task while the second part focuses on the verification of the requirements (Figure 1). The framework consists of four steps for the assessment of AIS. The first step is to estimate the risk for each dimension, discarding low-risk dimensions. For medium or high risk, each area within the dimension is evaluated. Next, objectives are defined based on the risk estimates and the means to achieve them are evaluated for sufficiency. Finally, all trustworthiness dimensions are verified across all dimensions.

Method toolboxes and evaluation frameworks help in ensuring transparent, explainable, and robust AIS. Industry companies such as IBM and Seldon have developed toolboxes such as AIX360 (Bellamy et al., 2018) and Alibi (Klaise et al. (2021), Van Looveren et al. (2019)) that include methods for explainability and uncertainty quantification. Other platforms such as Captum (Captum, 2023), Shapley (Lundberg and Lee, 2017), LIT (Tenney et al., 2020), and IBM Watson OpenScale (IBM, 2023) provide a variety of methods for model interpretability, fairness, bias, feature importance, and monitoring of AI models. Adversarial Robustness Toolbox (ART) is a framework for evaluating the adversarial robustness of neural networks, consisting of four types of attacks (Nicolae et al., 2018).

## 3. Method

The proposed framework starts by defining the application domain and identifying the stakeholders. An assessment of the risks associated with the entire application is then carried out in accordance with the EU directive. Specific objectives are set for different aspects of the AI application, covering the entire lifecycle from concept to operation and decommissioning. For each aspect, risks are derived from a high-level risk analysis and objectives are set to reduce them to an acceptable level. Means to achieve these objectives are defined, distinguishing between criteria and metrics ($C$) that refine the objective ($O$), and processes ($P$), documentation ($D$), and methods ($M$) to comply with the objective (Figure 2).

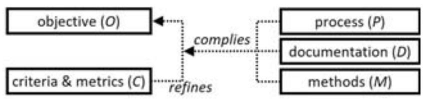



Fig. 2.: Dependency objectives and means.

The next step is to address different aspects of AI applications in more detail (figure 3), starting with data completeness and quality requirements.

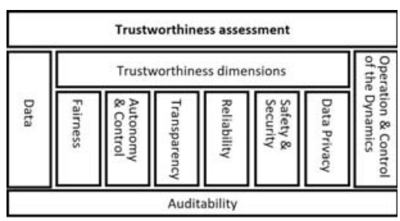



Fig. 3.: Extended trustworthiness dimensions.

Subsequently, the framework considers six dimensions as proposed by Poretschkin et al. (2021). This contribution focuses on the first four dimensions. For fairness, means are defined to ensure that AIS do not show bias or discriminate against individuals or groups. For autonomy and control, means are defined to ensure that AIS operate independently while allowing for human intervention and oversight. For transparency, means are defined to ensure that the decision-making processes of AIS are explainable and understandable. For reliability, means are defined to ensure that AIS operate as intended and produce consistent results.

Finally, the framework addresses issues that cut across all dimensions, including the control of dynamics during operation, such as changes in the domain, users, and models, and the need for procedures to be auditable. Lastly, an overall assessment is conducted to determine if the required reduction in overall risk has been accomplished.

## 3.1. *Data*

A dependable data set for a specific task requires careful consideration of four critical aspects: data quality, completeness, representativeness, and transparency. A comprehensive assessment of these aspects will result in a dependable data set that meets the objectives of each of the dimensions discussed in the following sections.

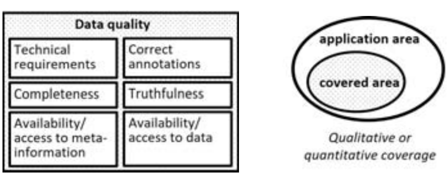
### 3.1.1. *Area: Data quality*

In this context, data quality is defined by two objectives:

(1) Achieve formal completeness and correctness of the used data set
(2) Establish a reliable database

Below the objectives in each section, we outline some of the means necessary to achieve them.

**Quality of the training, validation, and test data** is determined by qualitative and quantitative means for evaluating data quality (Figure 4(a)).




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

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

Technical means, such as data type, size, and format, are necessary for AIS use. Completeness and truthfulness ensure all necessary attributes and trustworthy data sources, respectively. Correct annotations are essential when algorithms label data. Relevance to the predefined task and accessibility of data/meta-information must also be assessed.

**Origin and quality of the data basis** considers the evaluation of the above defined requirements.

### 3.1.2. *Area: Data coverage*

To ensure data coverage of the application area (Figure 4(b)), two objectives must be met:

(1) Define the application area
(2) Ensure coverage of the application area

**Quantification of coverage** defines quantitative or qualitative metrics and intervals to assess how well the data covers the application area for a specific task. Quantitative metrics also refer to class and meta-information balance, while qualitative metrics include visualization and explorative analysis of high-dimensional data.

**Choice of data basis** includes documentation and justification of the used data. All the above requirements must also be met.

### 3.1.3. *Area: Representativity and Bias*

To prevent biased or unfair decisions made by the AIS, the following two objectives must be met in the training, validation, and test data:

(1) Provide bias-free training, validation, and test data
(2) Ensure fairness of training, validation, and test data

**Quantification of fairness in the training, validation, and test data** includes the documentation of quantitative metric(s) and appropriate intervals for the metric(s) to assess the fairness in the data (overview of fairness metrics in Figure 5).

**Quantification of bias (high similarity) in the training/validation/test data** includes the documentation of quantitative metric(s) and appropriate intervals for the metric(s) to assess bias in the data (e.g., cosine similarity based on raw data or latent representations).

**Verification of the unbiasedness of the data** refers to the method(s) for verifying the unbiasedness of the data (verification of the two criteria defined above).

### 3.1.4. *Area: Transparent data*

Transparent data refers to the (intrinsic) interpretability of all data.

**Transparent training and test data** ensures that users, those affected, and experts without prior knowledge can understand and interpret the data, including any preprocessing steps taken.

## 3.2. *Fairness*

Fairness in decision-making refers to the absence of prejudice or favoritism toward individuals or groups based on their inherent or acquired characteristics (Mehrabi et al., 2021). Biased decision-making processes can lead to algorithmic discrimination and unfair treatment of individuals or groups. Several types of fairness have been proposed in literature, including individual, subgroup, and group fairness (Dwork et al., 2011).

Only one area is addressed in this dimension, with two main objectives addressing fairness to individuals and (sub)groups.

(1) Ensure that the input is fair (see data)
(2) Ensure that the output is fair

**Identification of potentially disadvantaged groups** by assessing sensitive characteristics (e.g., gender, age, and ethnicity) present in the data for a specific task needs to be performed.

**Determination of an appropriate fairness concept** in the specific application context of the AIS has to be accomplished, including acceptable types of discrimination.

**Quantification of fairness** using measurable metrics (Figure 5) such as statistical measures, similarity-based measures, and causal inference needs to be established.

**Fair model building** must be achieved by documenting the model and learning process, and describing how fairness is promoted by the loss function and class/sample weighting. Fair adaptation and post-processing measures must be implemented to address any unfairness that may arise during or after the learning process.

**Testing the AI component on unseen data** (test and/or validation data) and documentation of the results is required for the intervals defined in the reliability dimension.

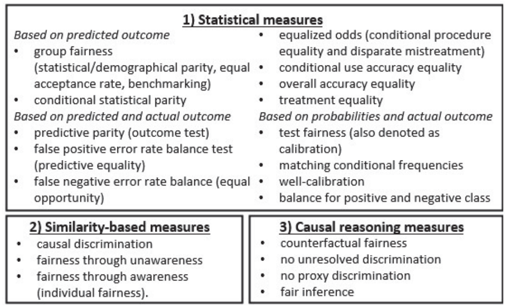



Fig. 5.: Fairness metrics overview.

**Test of the AI component on operational data** where the selected data must be documented and justified. The fairness of all relevant processing steps by the AIS integrated into the operation needs to be verified. Monitoring and documenting the fairness of the AIS during production is necessary to ensure its fair operation.

## 3.3. *Autonomy and Control*

The autonomy and control dimension is designed to address potential harm scenarios that may arise when autonomous AI components restrict users' or experts' perception or ability to act. The restriction of system autonomy when departing from the normal state is addressed in the safety dimension. To evaluate this dimension, the AIS must be classified into one of four categories:

**Human Control (HC):** The AI application is purely an assistance system. The human is involved in all decisions and initiates next steps based on the output of the AI application.

**Human-in-the-Loop (HIL):** The AI application acts partially autonomously but needs human operation/confirmation. Humans supervise, intervene, and correct AI decisions.

**Human-on-the-Loop (HOL):** Under normal conditions, the AI application is able to act autonomously without human intervention. The human mainly monitors the AI and is only involved as a decision maker in exceptional situations, where the human can override decisions made automatically by the AI application at any time.

**Human-out-of-the-Loop (HOOTL):** The AI application operates autonomously in all situations, including errors and unexpected events, completing tasks without human intervention. The user only decides whether to utilize the AI and sets up meta-commands (such as specifying the destination in an autonomous vehicle).

This dimension includes objectives such as human oversight and control mechanisms, human decision-making ability, and the comprehensibility of the AI component's decision-making process, leading to objectives related to transparency.

### 3.4. *Transparency*

The transparency dimension addresses potential damage scenarios caused by the lack of transparency in AIS, preventing safe and appropriate usage. According to Samek et al. (2019), transparency for AIS includes different types of transparency objectives for different stakeholders (Figure 6). Two areas are considered in the following: transparency for users and those affected and transparency for experts.

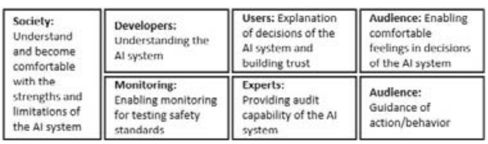



Fig. 6.: Transparency for different stakeholders.

#### 3.4.1. *Area: Transparency for users and those affected*

Two transparency objectives must be met:

(1) Define qualitative and quantitative criteria to assess the level of transparency.
(2) Achieve the appropriate level of transparency, in terms of interpretability and explainability.

**Assessment of explainability to users and those affected:** Selection and justification of qualitative criteria, like unambiguousness and comprehensibility, and quantitative criteria (metrics) for assessing the explanation methods for the AIS (e.g., Chan et al. (2022).

**Interpretability of the ML model:** The suitability of an intrinsic-interpretable model for the AIS should be assessed.

**Comprehensibility of the functionality of the ML model:** This should be ensured via proper documentation and visualization of the ML model (schematic diagram of the architecture etc.).

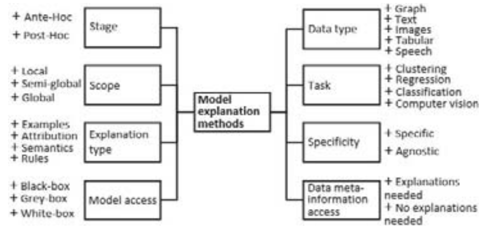



Fig. 7.: Taxonomy for categorizing transparency methods based on different criteria (adapted Ding et al. (2022)).

**Selection of explanation methods for the obtained results:** This selection should be justified and documented, considering the qualification of the users and those affected. It should be based on the adapted taxonomy by Ding et al. (2022) in Figure 7 for each specific task of the AIS.

As can be seen in Figure 7, model explanation can be further divided into explaining an ML model on a local and global level. Local model explanations aim to explain the reasoning behind the prediction of the ML model for a specific input, while global explanations explain the model's overall behavior and input-output relationship (see Figure 8).

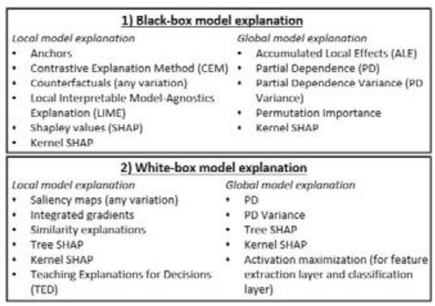



Fig. 8.: Model explanation methods overview.

Additionally, model explanations should undergo statistical and human evaluations and a process for responding to user requests should be established.

#### 3.4.2. *Area: Transparency for experts*

Transparency for experts has similarities with the previous area. However, the focus is on validation and on the technical traceability and reproducibility of outputs of the AI application by experts. The technical level is correspondingly higher.

The main objectives are:

(1) Apply introspective explanatory methods for transparent and understandable decisions.
(2) Assess the inherent "logic" of the AI.
(3) Identify potential causes of errors and systematic model weaknesses.

When defining the requirements for the properties of the explanation methods, following aspects should be considered: Scope, design, degree of transparency, depth of introspection in relation to the model outputs, time frame, complexity, and stability of the methods.

### 3.5. *Reliability*

Reliability refers to the ability of an AIS to consistently perform its intended functions, while robustness refers to the ability to maintain performance and functionality in the presence of disturbances. The input space can be divided into regular, irregular, and error cases (Figure 9), where the system must be both reliable and robust to handle minor disturbances in the regular case and major disturbances in the irregular case. However, in an error case, where the data is outside the application area, the system may not be able to handle it, leading to potential errors.

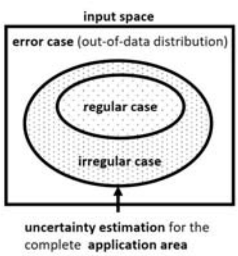



Fig. 9.: Visualization of the different regions of the complete input space.

#### 3.5.1. *Area: Reliability in the regular case*

To ensure reliability in the regular case, several objectives must be met:

(1) Ensure that the data used to develop the AI application covers the range of inputs expected during operation (see data).
(2) Quantify and evaluate the performance of the AIS using suitable metrics.
(3) Mitigate the risk of errors and misjudgments.

**Quantification of reliability** of AI applications is established using mathematical and statistical metrics, including performance metrics and loss functions.

**Quantification of the coverage of the application area** is to be done by defining and justifying target intervals for the coverage measure, as stated in the data objectives. To enhance the training data's coverage various methods can be used to generate additional input data, as depicted in Figure 10.

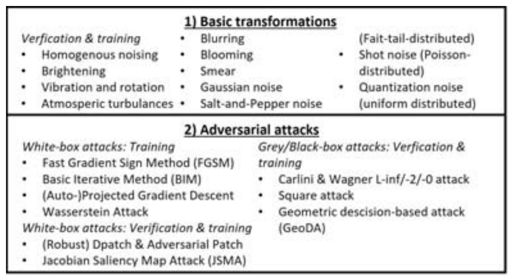



Fig. 10.: Common reliability methods overview.

**Choice of AI component:** Documentation should exist that explains how the chosen model components (training algorithm, loss function, etc.) are related to the reliability requirements.

**Systematic search for vulnerabilities:** One important aspect of ensuring reliability is systematically searching for vulnerabilities. This can be done through techniques such as closed-loop testing or introspective explanation methods, and any weaknesses that are discovered should be documented and addressed with appropriate measures.

#### 3.5.2. *Area: Robustness*

In this area, three objectives must be met:

(1) Define application boundary
(2) Strengthen AI robustness
(3) Detect and intercept errors

For the objectives, an appropriate set of methods is to be selected from a list of methods based on the taxonomy proposed in Figure 11.

#### 3.5.3. *Area: Uncertainty Estimation*

To ensure that the AI application provides an accurate statement of confidence in its results, two main objectives are required:

(1) Determine an uncertainty metric
(2) Select an appropriate uncertainty assessment method

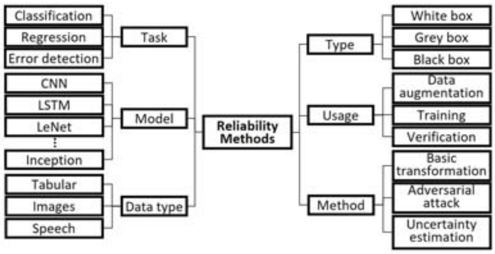



Fig. 11.: Taxonomy of reliability methods.

### 3.6. *Safety and Security*

Functional safety and IT security are assessed in this dimension. Functional safety involves designing and implementing an AIS to minimize harm to people or the environment, including providing corrective mechanisms for unexpected behavior. Meanwhile, IT security ensures the integrity and availability of the system by protecting against unauthorized access, modification, and destruction, protecting the system from cyber-attacks, and ensuring availability.

### 3.7. *Data Privacy*

The data privacy dimension aims to identify and document data protection risks in AI applications, accounting for AI-specific challenges, to assist data protection officers in decision-making. This includes non-compliant use of personal data, the risk of re-identification of individuals in a data set, unwanted disclosure of business-relevant information by the AI application, and risks from changing data processing requirements.

### 3.8. *Control of Dynamics*

This dimension addresses the potential consequences of a model and concept drift and their influences on the remaining trustworthiness dimensions. A model drift can occur if the model is further trained on new incoming training data during the operation phase. In contrast to model drift, concept drift is a result of changed external conditions that lead to new requirements for the AIS. Those can be triggered by changes in legislation, social values, or also hardware.

### 3.9. *Auditability*

Auditability refers to the ability to audit the technical documentation of an AI application, including its development, functionality, and training data. This involves specifying which parts of the AI application require documentation and to what extent, as well as the level of traceability and reproducibility needed for the outputs. Traceability in AI involves tracing the history and derivation of an AIS's decision (e.g., logging inputs, predictions, explanations, and newly captured data, storing relevant parameters). Reproducibility, on the other hand, involves the ability to replicate an AIS's results using the same data and algorithms (e.g., saving random seeds, documenting hardware specifications and each task).

## 4. Use Case

The framework was tested on an assistive AI application for classifying skin lesions using the ISIC Archive (2019) data set comprising 25,331 images categorized into two classes: 20,181 benign and 5,150 malignant samples. The dermatologist captures an image of the skin lesion, which is evaluated by the ML model. The dermatologist determines the appropriate treatment based on the model's assessment and additional interpretive information (Figure 12).

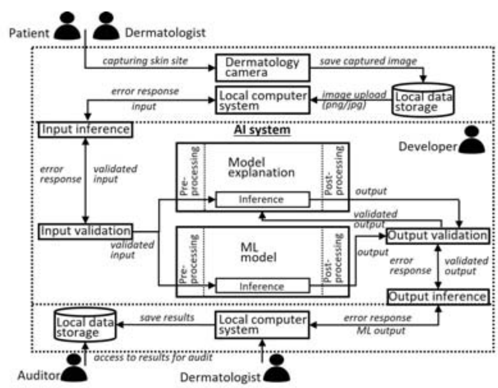



Fig. 12.: Representation of the use case for the assistive of skin lesion detection application.

Based on the overall risk assessment, the relevant overall aspects, individual dimensions, and areas are identified, which subsequently determine the objectives (including associated criteria).

To meet these requirements, specific measures were defined, as exemplified in Figure 13, to satisfy the requirements and their criteria.

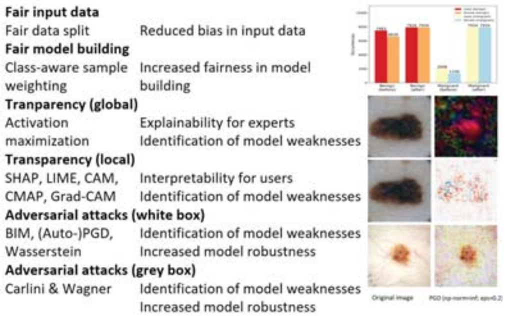



Fig. 13.: Exemplary methods and their objectives within this use case.

## 5. Conclusion

The results confirm the usefulness of the framework. However, successful implementation requires careful selection of measures and methodologies aligned with the dimension's requirements and the specific application. Merely evaluating the process is insufficient; a comprehensive technical assessment of the AI application at multiple dimensions is essential. By separating certain aspects, such as data, from individual dimensions, duplicate requirements can be avoided. In addition, conducting cross-dimensional assessments at appropriate intervals ensures the comprehensiveness of the framework.

## References

Bellamy, R. K. E. et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv:1810.01943*.

Captum (2023). Model interpretability for pytorch. *https://captum.ai/*.

Chan, C. S. et al. (2022). A comparative study of faithfulness metrics for model interpretability methods. *arXiv:2204.05514*.

Council of European Union (2021). Laying down harmonized rules on artificial intelligence com(2021)206 final.

DIN, DKE (2023). Artificial intelligence standardization roadmap. *https://www.dke.de*.

Ding, W. et al. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences 615*, 238–292.

Dwork, C. et al. (2011). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

IBM (2023). IBM Watson Studio: build trust in AI. *https://www.ibm.com/cloud/watson-studio*.

IEEE (2022). IEEE CertifAIEd: the mark of AI ethics. *https://engagestandards.ieee.org*.

ISIC Archive (2019). International skin imaging collaboration: Melanoma project website. *URL https://isic-archive.com*.

Klaise, J., A. V. Looveren, G. Vacanti, and A. Coca (2021). Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research 22*(181), 1–7.

LNE (2023). Certification of processes for AI. *https://www.lne.fr*.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pp. 4765–4774.

Mehrabi, N. et al. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv. 54*(6).

Nicolae, M.-I. et al. (2018). Adversarial robustness toolbox v1.0.0. *arXiv:1807.01069*.

Poretschkin, M. et al. (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog).

Samek, W. et al. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer.

Soudain, G. (2021). First usable guidance for Level 1 machine learning applications: A deliverable of the EASA AI Roadmap.

Tenney, I. et al. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. *arXiv:2008.05122*.

Van Looveren, A. et al. (2019). Alibi detect: Algorithms for outlier, adversarial and drift detection. *https://github.com/SeldonIO/alibi-detect*.