Benchmarking Trust: A Metric for Trustworthy Machine Learning*

Jérôme Rutinowski$^{1,2[0000-0001-6907-9296]}$, Simon Klüttermann$^{1,2[0000-0001-9698-4339]}$, Jan Endendyk^{1}, Christopher Reining$^{1[0000-0003-4915-4070]}$, and Emmanuel Müller$^{1,2,3[0000-0002-5409-6875]}$

[1] TU Dortmund University, Dortmund, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
[3] Research Center Trustworthy Data Science and Security, Germany
jerome.rutinowski@tu-dortmunde.de

Abstract. In the evolving landscape of machine learning research, the concept of trustworthiness receives critical consideration, both concerning data and models. However, the lack of a universally agreed upon definition of the very concept of trustworthiness presents a considerable challenge. The lack of such a definition impedes meaningful exchange and comparison of results when it comes to assessing trust. To make matters worse, coming up with a quantifiable metric is currently hardly possible. In consequence, the machine learning community cannot operationalize the term, beyond its current state as a hardly graspable concept.

This contribution is the first to propose a metric assessing the trustworthiness of machine learning models and datasets. Our FRIES Trust Score is grounded in five key aspects we understand to be the fundamental building blocks of trust in machine learning – fairness, robustness, integrity, explainability, and safety. We evaluate our metric across three datasets and three models, probing the metric's reliability by enlisting the expertise of ten machine learning researchers in its application. The results underline the usefulness and reliability of our method, seeing distinct overlaps between the participants' ratings.

Keywords: Trustworthiness · Machine Learning · Operationalization

* This research has partly been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

1 Introduction

Machine learning is becoming an increasingly integral part of computer science. In addition, it is finding its way into interdisciplinary research. For many interdisciplinary applications, the data used for machine learning models is gathered all across real life and the internet, often without users even being aware of it. The ubiquity of data driven models and the thriving interest in the generation of the data needed to train them, is also accompanied by skepticism and doubts. These worries are further amplified by the fact that the notoriously intransparent deep learning models [5] have become the state of the art for many challenges. Thus, aspects such as the reliability of their reasoning process, their sensitivity to unforeseen circumstances, or the privacy considerations regarding the data they are trained on, are a concern of many researchers and users alike. It is therefore becoming increasingly important to evaluate the trustworthiness of machine learning models and datasets.

However, what exactly constitutes the term 'trust' and what aspects it encompasses, is not properly defined in the current literature. This is a crucial limitation, as without such a definition, the concept cannot be quantified and thus operationalized. In the past, a dataset like DukeMTMC [40], which was the state of the art for person re-identification tasks for a significant period of time, was used and accepted by the research community, without a consideration for, e.g., data privacy concerns. The dataset was recorded at Duke University's campus in 2014. It was an outstanding dataset in terms of size, amount of cameras and data diversity. However, the dataset and its associated publication have since been retracted [37], as the recordings were captured and published without the knowledge of the recorded individuals. This renders DukeMTMC an excellent example of a dataset containing sensitive data related to privacy rights and thus plays a role in considerations for secure and fair utilization. Does this make the DukeMTMC dataset and any machine learning application trained on it untrustworthy? Does this make other datasets more trustworthy in comparison? And if so, how would we know? Is data privacy and fair use an aspect of trustworthy machine learning? And if so, how would we measure it? Currently, these questions cannot be answered properly, hence motivating this work.

This contribution's goal is to provide a metric that permits the quantification of the overall trustworthiness of machine learning models or datasets, composed of the key aspects that make up trust. To this end, we explore aspects of trust that are frequently mentioned in literature. These aspects are incorporated into an overall key figure by virtue of a trust score, inspired by the failure mode and effects analysis [44]. We evaluate this concept of trust through the application of our metric on three distinct datasets and three distinct models. In addition, the reliability of the resulting metric and definitions is evaluated by comparing the results stemming from different users when applying the metric to the same models and datasets.

The structure of the paper is as follows: The following section will explore and present the key aspects of trustworthy machine learning to formulate operational definitions to be used for our contribution. Section 3 will detail the methods with which we intend to quantify trust and subsequently evaluate the resulting metric. In Section 4, the metric will then be applied to three datasets and models, providing the reader with a first insight into what the use of the metric would entail. In addition, the agreement of different users when faced with the same dataset or model and thus our metric's reliability is evaluated. Finally, in Section 5, we will evaluate the validity of our findings and discuss their current limitations.

2 Aspects of Trustworthy Machine Learning

As this contribution aims to provide a metric for the quantification of trust, we first need to operationalize the underlying aspects of the notion of trust. We did so through a deductive category formation [30] in accordance with the relevant literature. We opted for this approach, because while there may be a plethora of discussion about the aspects of trust available, they are not properly structured yet. Other researchers have proposed other sets of aspects of trustworthy machine learning, as did [48], who understand fairness, explainability, auditability, and safety to be the main aspects of it. Alternatively, [47] claim fairness, privacy, security, and integrity to be the central parts of trustworthy machine learning. The authors, however, do not quantify these aspects and given our research, without proper definitions and metrics and the inclusion of the aspect of robustness, this is incomplete.

This section thus provides an overview of the key aspects of trustworthy machine learning encountered in literature, namely fairness, robustness, integrity, explainability, and safety, that are most prominent in the relevant literature [25,42,50,51]. Each subsection is dedicated to one of these aspects and outlines the diverse range of views that researchers express, in part conflicting and in part corroborating each other's perspectives. The corresponding literature has been found by combining the five aspects with the search strings 'machine learning', 'artificial intelligence', 'data science', and 'dataset'. The relevant contributions have been sorted by year of publication to provide an up-to-date overview. Also, when possible, we identified the first-ever mention of each aspect to display the shift in perception. Each subsection summarizes relevant research on the respective aspect in a table, stating whether an explicit definition or metric was proposed in the respective paper. Several papers provided similar notions, so that listing all of them would reduce comprehensibility without adding value. In these cases, we only refer to the paper that has yielded the largest impact on the research community. Each subsection concludes with an operational definition of the respective aspect of trust, which is subsequently used throughout this contribution.

2.1 Fairness

When it comes to requirements in the context of evaluating machine learning models and datasets, fairness is a topic of interest frequently encountered, yet not defined in a standardized manner and instead represented through multiple notions that are in part incompatible with one another [50]. Fairness in decision-making is a fundamental concept rooted in the principles of equal treatment and the avoidance of disparities and discrimination [11]. Alternatively put, it is the attempt to use ground-truth values, that are unbiased [42]. Yet it is considered a challenging concept to measure [48], in particular in the domain of machine learning [7]. Some approaches for fairness metrics have been proposed [42], albeit ones that are very specific for their use case. This is because the concept's meaning strongly depends on the context it is used in: while fairness might not be relevant for models or data used in production plants, it tends to be crucial for applications that support critical decisions, e.g., credit scoring or rating of job applicants [42].

Table 1: An overview of relevant literature on fairness in machine learning.

Ref.	**Year**	**Author(s)**	**Notion of fairness**	**Met.**	**Def.**	**Cit.**
[42]	2022	Schmitz et al.	Use of unbiased ground-truths	Yes	No	5
[47]	2022	Thuraisingham	Unbiased decisions, exclusion of variables such as gender	No	No	9
[28]	2022	Strobel et al.	No unequal treatment of demographic groups due to inequalities in training data	No	No	10
[46]	2021	Thiebes et al.	Amend past inequities, ensure equal distribution of benefits, reduce new harms and inequities	No	Yes	394
[48]	2020	Toreini et al.	Providing non-discriminatory and unbiased outcomes	No	No	233
[7]	2020	Caton et al.	Parity in outcome distributions, proportional attributions, treatment equality	Yes	Yes	359
[11]	2019	Dodge et al.	No difference in treatment of individuals that differ only in a sensitive attribute	No	No	289
[23]	2017	Kusner et al.	Decisions remaining the same in both the real scenario and a counterfactual scenario, reversing an individuals demographic group	Yes	Yes	1579
[18]	2016	Hardt et al.	Protected attributes are not to be considered during decision-making	Yes	Yes	3810
[53]	2013	Zemel et al.	Exclusion of information, that could imply an individual's group adherence (fairness through ignorance)	No	Yes	1786
[32]	1980	Mitchell	First paper encountered on biases in machine learning: Understanding of biased decisions impacting outcomes	No	No	289

The notion of individual fairness revolves around ensuring equitable outcomes for individuals with similar attributes [23, 53]. Another approach, in this context, is to entirely exclude immutable attributes that could lead to any such discrimination [47], although the opposite idea is proposed by other researchers as well [23].

Conversely, group fairness emphasizes the importance of mitigating biases related to specific attributes [7,11,28] and to ensure that sub-groups are not being treated differently than the main group [18]. The relevant group attributes, often demographic in nature [7, 23], encompass factors such as ethnicity and gender. [28] warn that group fairness might come at the cost of privacy, since sensitive attributes might need to be revealed for this purpose. However, achieving group fairness does not mandate identical representation across all categories. Rather, it entails that the relative distribution of outcomes aligns with the distribution of attributes in the original dataset [53].

Based on our research (see Tab. 1), for the purpose of a first attempt of operationalization, we propose the following operational definition:
Fairness in machine learning ensures the equal treatment of both groups and individuals, and the prevention of biases, promoting equitable outcomes.

2.2 Robustness

Robustness, sometimes referred to as reliability, is concerned with the maintenance of the expected model performance, even under strong scrutiny, such as exceptional, manipulated, noisy or unseen data [12, 15, 25, 28].

[12] state that robustness is an ’overloaded’ term that allows for a multitude of interpretations. The authors claim to have analyzed 53 studies, that by and large do not define what their understanding of robustness is. Thus, these interpretations can range from mere task performance on hold-out datasets to maintaining this performance on manipulated inputs (i.e., noise or adversarial attacks). The change in performance on edge cases, exceptions, or corrupted data, can also be perceived as a part of robustness.

According to [28], robustness can be understood as a model’s resistance to adversarial attacks. A robust model should not just correctly predict the label for ordinary, expected inputs but also remain correct if the input is manipulated [25]. In this sense, a models output should not be overly sensitive to a change in inputs [50]. In addition, the authors mention the risk of crowd-sourced data. The data quality might suffer due to the data collection process being noisy or the crowd workers performing sub-par work or even acting maliciously. [28] state that most modern machine learning models are not robust against such threats. To mitigate these risks, the reduction of a model’s sensitivity to changes of minor parts of the training data or the approach to entirely ignore outliers during training, have been proposed [28]. [52] propose the use of graph models as particularly robust model types. Concerned with robustness in healthcare imagery, [9] stress the importance of a model’s robustness against small perturbations and distribution shifts. One solution to this is provided by [38], who provide the Python package

Foolbox, which allows users to apply different adversarial attack types to their model's data.

Table 2: An overview of relevant literature on robustness in machine learning.

Ref.	Year	Author(s)	Notion of robustness	Met.	Def.	Cit.
[25]	2023	Li et al.	The ability to deal with execution errors, erroneous inputs, or unseen data	Yes	Yes	152
[28]	2022	Strobel et al.	Correct predictions not only for ordinary inputs but also for manipulated or noisy data	No	No	10
[50]	2021	Wing	How sensitive the system's outcome is to a change in the input	No	Yes	130
[12]	2021	Drenkow et al.	Maintaining task performance on manipulated or modified data as well as cross-domain generalizability	Yes	Yes	49
[52]	2021	Xu et al.	Maintaining results despite noise or adversarial attacks	Yes	No	15
[9]	2021	Darestani et al.	Ensuring correct predictions despite adversarial perturbations or distribution shifts	Yes	No	61
[15]	2018	Goodfellow et al.	Providing accurate predictions even in exceptional cases and for modified inputs	No	No	396
[38]	2017	Rauber et al.	A model's resistance to data perturbations	Yes	Yes	619
[3]	1976	Ashton	First paper encountered on model robustness: maintaining prediction accuracy when encountering changes in the information environment or parameters of the model	No	No	14

In summary, the lack of definitions and the variety of notions hardly allow for the quantification or comparison of machine learning approaches' robustness. In this contribution, based on Tab. 2, we propose the following operational definition:
Robustness in machine learning denotes the capability to sustain an expected performance despite encountering exceptional, manipulated, or noisy data.

2.3 Integrity

Integrity may describe multiple concepts in the realm of machine learning, such as the accuracy of the data or the model used as well as overall procedural intactness [47]. Influences such as the corruption of data or the involuntary change in model architecture may have an impact as well [47]. [19] stress the importance of model integrity, especially when deployed remotely, where changes to it could be made without authorized users noticing. In the same vein, [46] recommend the use of distributed ledger technology, to both prevent unwanted changes in a system as well as ensure the traceability of changes that were made. Besides the prevention of such tampering efforts, [48] name the detection and subsequent repair of tampered data (i.e., a reversal of the changes made) as an important mitigation step. Thus, ensuring data integrity is understood by some researchers as noticing and tracking changes [31].

Table 3: An overview of relevant literature on integrity in machine learning.

Ref.	Year	Author(s)	Notion of robustness	Met.	Def.	Cit.
[47]	2022	Thuraisingham et al.	Accuracy of the data used, procedural coherence, algorithmic correctness	No	No	9
[21]	2022	Hou et al.	Preventing malicious changes of the training data, that are meant to tamper with the model's predictions	No	No	19
[24]	2022	Kuttichira et al.	Preventing model tampering by comparing predictions of original and potentially compromised model	Yes	No	11
[48]	2020	Toreini et al.	Prevention of tampering as well as detection and repair of tampered data	No	No	233
[31]	2020	Meske et al.	Changes to data should not be able to occur unnoticed and should be traceable	No	No	66
[19]	2018	He et al.	Verifying that the model's architecture has not been tampered with when deploying it remotely	Yes	No	25
[17]	1991	Haber et al.	First paper encountered on data integrity: an approach of how digital documents can be time-stamped to ensure integrity	No	No	585

Traditionally, integrity-proofing measures of such kind would entail some form of hashing procedure. This, however, can hardly be applied in the case of machine learning, as access to the model itself is often restricted [24]. [24] therefore propose a Bayesian Compromise Detection (BCD) algorithm that aims to maximize the difference in the prediction of an original model and compares it to a potentially compromised model. While this is not a metric per se and the authors did not formally define their understanding of the term integrity in this context, being able to measure such differences through their proposed algorithm is beneficial in its own right.

Similarly to this approach, [21] propose a similarity-based integrity protection method for deep learning systems (IPDLS). This method is based on an anomaly detection approach that measures the similarity between suspicious samples and samples in a preset verification set. Again, this enables the assurance of integrity, providing a step in the right direction for the research community.

Based on our research, essentially resumed in Tab. 3, we propose the following operational definition:
Integrity in machine learning is the effort of preventing untraced or unauthorized changes to a data or a model, mitigating the risks of tampering efforts and reverting them if needed.

2.4 Explainability

In the realm of machine learning, explainability has emerged as a crucial factor of user trust, ethical considerations, and the widespread adoption of AI technologies [6]. Nevertheless, the concept remains vague in nature. Researchers agree that its use is quite ambiguous and that the term is often used synonymously with other terms such as transparency, intelligibility, comprehensibility, and interpretability [1, 6, 16, 42, 48].

Researchers have proposed taxonomies of explainability, e.g., concerning the understanding of a model's workings (pre-hoc) or merely its decisions (post-hoc) [20]. When analyzing pre-hoc explainability, [1,28] refer to the model's semantics being in alignment with the respective task semantics. In contrast, [6, 25, 48, 50] understand the concept as the provision of reasons for the decisions (i.e., predictions) that have been made by a given model. [20] stress the importance of determining who the users in question are, as this determines the way in which explainability is shaped. The authors also run an experiment in which they ask users to rank the explainability of a set of models on a Likert scale, aiming to provide a first way of evaluating a model's explainability.

Some works focus primarily on the notion of explainability concerning human understanding [1, 16, 29, 34, 48]. This task can vary depending on the respective model and the opacity of certain model types (e.g., DNNs versus SVMs) is discussed in this context [29]. Little research has been encountered concerning the explainability of data. [16] discuss data in the context of human understandability, along with models and their predictions. The authors understand images and text to be more explainable than tabular or vector/matrix data, providing users with a more intuitive access to data interpretation.

In summary, these studies paint a comprehensive picture of the evolving landscape of explainability in machine learning, emphasizing its multifaceted nature as well as its role in fostering trust and user comprehension. Based on our research (see Tab. 4) we propose the following operational definition:
Explainability in machine learning is the systematic effort to render decision-making procedures of models interpretable and datasets understandable, facilitating both the insight into their inner workings and aiding stakeholders in validating outputs.

Table 4: An overview of relevant literature on explainability in machine learning.

Ref.	Year	Author(s)	Notion of explainability	Met.	Def.	Cit.
[20]	2023	Herm et al.	Overview of How, Why, Why-Not, How-To, and What-Else approaches to explainability	No	Yes	46
[25]	2023	Li et al.	The ability to understand how a model made its decision	No	Yes	152
[42]	2022	Schmitz et al.	Technical transparency of artificial intelligence	No	Yes	5
[28]	2022	Strobel et al.	Model semantics in alignment with task semantics	No	No	10
[6]	2021	Burkart et al.	Providing reasons for decisions made	No	No	607
[50]	2021	Wing et al.	The ability to justify the model's outcome with an explanation that a human can understand	No	Yes	130
[48]	2020	Toreini et al.	Explaining and interpreting the outcome of a decision to stakeholders in a humane manner	No	No	233
[1]	2020	Arrieta et al.	Providing a model's details and reasoning to make its functioning clear or easy to understand	No	Yes	5279
[29]	2020	Marcinkevics et al.	Decisions that can be comprehended by a human and the explanation of predictions made by opaque models	Yes	No	115
[34]	2018	Montavon et al.	Mapping abstract concepts that produced a decision into a domain that the human can make sense of	No	Yes	2462
[16]	2018	Guidotti et al.	The extent to which models, data or predictions are human understandable	No	Yes	3837
[4]	1995	Auer et al.	First paper encountered on explainable models: focusing on the drawbacks of explainable decisions and performance	No	No	170

2.5 Safety

Safety relates to the security of models and data, and the associated protection of privacy, as defined in Art. 4 GDPR [13]. Recently, the European Union has ratified the Artificial Intelligence Act, which encompasses safety as one of its concerns regarding the use of machine learning [14].

The term safety is often used synonymously with terms such as security, privacy, and dependability [22, 28, 33, 47]. Similar to fairness, a distinction can be made throughout the process between data that can concern either individuals or groups of people [28].

In this context, [33] offer a checklist of safety measures and strategies, which could be suited as part of a metric. [43] focus on the term of privacy and the risks that come along with data usage in machine learning. The authors name membership inference attacks as a threat to privacy, since these attacks aim to determine whether a given data point was used to train the model or not. By doing so, information about an individual can be inferred, e.g., when it becomes known that they are part of a group of individuals used in a health-care related dataset (e.g., individuals suffering from an illness, they may not want to have disclosed). The first work that could be found that is concerned with the issue of inferring confidential information through an indirect manner was [8], coining the term statistical disclosure.

Table 5: An overview of relevant literature on safety in machine learning.

Ref.	Year	Author(s)	Notion of safety	Met.	Def.	Cit.
[33]	2022	Mohseni et al.	A concept to protect from non-desirable outcomes, like data theft and privacy violations	No	Yes	21
[42]	2022	Schmitz et al.	Ensuring the proper functioning of a system while safeguarding it against potential vulnerabilities	No	No	5
[28]	2022	Strobel et al.	Ensuring the containment of information about individual data records beyond general patterns	No	Yes	10
[47]	2022	Thuraisingham et al.	Models only accessing the data they are authorized to, in order to carry out designated task; Data ensuring high prediction performance while maintaining privacy	No	Yes	9
[43]	2021	Song et al.	Ensuring anonymity of individuals whose information is part of datasets	No	No	257
[10]	2020	Decristofaro	The property that a model's output does not differ significantly for two versions of the data differing by only one individual data point	No	Yes	80
[36]	2018	Papernot et al.	Confidentiality of a model's architecture and parameters, and of the data sources	No	No	529
[22]	2014	Zhanglong et al.	The protection of private data from leakage, especially the information of individuals	No	Yes	311
[8]	1977	Dalenius	First paper encountered on dataset safety: statistical disclosure of an individual's information as a risk	No	No	513

Analogous to this issue, [10] introduce the concept of differential privacy, which is concerned with a model's output differing when removing only one data point, again potentially revealing confidential information about an individual. Other research focuses on the confidentiality of the models themselves, i.e., the model's architecture and specific parameters, that might be proprietary and thus confidential by nature [36]. The exposure of such intellectual property can be a safety threat in and of itself but can also impact the privacy of the data source used, especially when the users of the model are not trusted individuals (i.e., in the case of models that are publicly available).

In summary, these studies illuminate the complex facets of safety in machine learning, safeguarding sensitive information by employing privacy-preserving techniques. Based on our research (see Tab. 5) we propose the following operational definition:

Safety in machine learning encompasses the protection of confidential or proprietary model architectures and parameters as well as data from unauthorized access.

3 Measures of Quantification and Operationalization

Despite the intuitive notions of trust most of us have, sociological definitions vary wildly and are inconsistent as to whether trust is an attitude, an intention, or a behavior [2]. Based on the research presented in the preceding section, we propose a first set of definitions of what we perceive to be the overarching themes and thus key aspects of trust in machine learning. Having identified and defined these key aspects, we can now begin to articulate the concept of trust itself. However, what is still needed is a manner of quantifying the aspects for which we have formulated these definitions and subsequently, a manner to evaluate the chosen quantification approach. Thus, in this section, we will explore a method for quantifying trust in the FRIES Trust Score. Beyond that, we assess the reliability of the developed metric. Human experts will be asked to assess machine learning models and datasets using our metric. This is necessary, because we expect that the individual evaluations of the same datasets and models using our metric might not entirely coincide. Thus, a strong overlap in the resulting scores will be understood as an indication that the terms our score refers to are well articulated and possess a certain degree of reliability. If the inverse is the case, this would imply a certain ambiguity in the procedure, which could hinder the reproducibility of the results.

Finally, based on the research presented in the preceding section, we propose the following operational definition of the concept of trust in machine learning:

The concept of trust in machine learning comprises the fair use of data, robust performance when encountering anomalous data, the assurance of data and model integrity, the provision of explainable decisions as well as the safe use of confidential information.

3.1 Quantifying the Notion of Trust

As we observe significant similarities between quality assurance procedures and the evaluation of concepts such as trust, our research prompted us to examine the models employed in quality assurance practices. Other researchers of trustworthy machine learning, such as [42] have done so before, proposing their AI risk scheme based on this. A model that is prominently used in the context of quality assurance is the Failure Mode and Effects Analysis (FMEA), often also referred to as risk mode and effects analysis. FMEA is a systematic and semi-qualitative risk analysis method [44]. Its use is recommended in many standards, such as in quality management [26], and depending on the area of its application, FMEA is used to assess, e.g., systems, software or processes [26, 44].

Table 6: Risks for trustworthy machine learning models (gray lines) and datasets (white lines) that can be chosen by users.

Aspect	Risk
Fairness	Decisions made by the model are biased against certain groups or individuals
	User inputs are requested in a biased manner
	Performance differs for certain groups or can only be applied to certain groups
	The dataset is not representative of the application (sampling bias)
	The dataset includes protected attributes
	The dataset perpetuates biases (e.g., is generated from unfiltered web data)
Explainability	The model's decision-making process is not transparent
	The model's architecture is unknown or prohibits its interpretation
	Stakeholders cannot validate the model's outputs
	No documentation of the data collection and annotation process
	The dataset is not human understandable
	Lack of clarity on how missing values or outliers are handled in the dataset
Safety	Decisions or internal representations could reveal sensitive information
	Insufficient access control to proprietary model
	Erroneous decisions might lead to critical consequences
	Insufficient access control to proprietary data
	Exposure of sensitive information through metadata or auxiliary data
	Lack of transparent data governance policies (e.g., data usage agreements)
Robustness	Risk of adversarial or inversion attacks not mitigated
	The model does not generalize to different datasets
	Repeated model executions do not generate the same or similar outputs
	The dataset does not contain edge cases or outliers
	The data is susceptible to distribution shifts
	The data contains harmful anomalies or perturbations
Integrity	It cannot be guaranteed, that the model was not tampered with
	No output uncertainties are given
	Changes made to the model cannot be tracked
	It cannot be guaranteed, that the data was not tampered with
	Changes made to the data cannot be tracked
	Pronounced labeling uncertainties cannot be ruled out

Table 7: Scales for the probability of the Occurrence (O), the Significance (S), and the probability of the Detection (D) of a risk for trustworthy machine learning.

Occurrence (O)		**Significance (S)**		**Detection (D)**	
Probability		Impact		Probability	
Impossible	10	Negligible	10	Certain	10
Unlikely	9	Barely perceptible	9	High	9
Very low	7-8	Insignificant	7-8	Moderate	7-8
Low	4-6	Moderate	4-6	Low	4-6
Moderate	2-3	Severe	2-3	Very low	2-3
High	1	Extremely severe	1	Unlikely	1
Certain	0	Unacceptable	0	Impossible	0

At the core of FMEA lies the preventive analysis of possible faults and the associated causes of a subsequent failure. Per failure type, users are asked to list potential faults and their respective consequences and causes. For each failure, the probability of occurrence (O), the significance of the failure (S) and the probability of detection (D) are evaluated on a scale of 1 to 10 (higher numbers representing an increased probability of occurrence, a higher significance, and a lower probability of detection). These resulting values are multiplied with each other and provide a risk priority number (RPN) between 1 and 1000 for each failure, whereby a higher risk priority number represents a higher risk [44]. For these individual RPN, a measure of risk mitigation is being proposed and upon implementation a new RPN per failure can be calculated. Thus, no overall score is calculated and the focus lies on the suggestion of quality improvement measures.

Table 8: The calculation of the FRIES Trust Score in the form of a table.

Aspect	Risk	O	S	D	Π	$\bar{\Pi}$	ω	T_ω
Fairness	Inputs requested in a biased manner	4	4	8	5.04	5.04	0.2	1.01
Robustness	Risk of model inversion attacks	4	8	9	6.6	5.89	0.2	1.18
	Risk of adversarial attacks	7	4	5	5.19			
Integrity	The model is not open source	3	9	2	3.78	3.78	0.2	0.76
Explainability	Illusion of Explanatory Depth	8	4	5	5.43	5.43	0.3	1.63
Safety	Decisions reveal sensitive information	6	3	6	4.76	4.76	0.1	0.48
							T	5.06

For the task at hand, FMEA is particularly well suited, compared to other quality assessment methods, such as the utility value analysis. This is due to its differentiated OSD evaluation of failures, which permits qualitative evaluations to be translated into quantitative values, if so desired. However, some adaptations to FMEA are necessary, eventually developing our own approach based on it. In a first step, we observe failures as parts of the previously defined aspects of trustworthiness, i.e., fairness, robustness, integrity, explainability, and safety. Instead of potential failures, we observe potential limitations of these defined aspects, i.e., phenomena that could jeopardize the respective aspect and we consider to be risks. For this, we provide the users with a table of limitations/risks per aspect, that can be applied to either datasets or models (see Tab. 6) and of which one to three can be chosen per aspect. Based on this, the qualitative assessment is now translated into a quantitative assessment by virtue of an adapted OSD approach. In contrast to the standard FMEA approach, the selectable value of 0 was added, for situations in which the probability of occurrence is certain, the significance of a risk entails unacceptable impacts, or its detection is impossible (see Tab. 7).

Algorithm 1 FRIES Trust Score T calculated with our novel approach.

```
Require: $\omega_i\ \forall i \in [0,5);\ \omega_i \geq 0.1$   ▷ Set importance for each of the five aspects
Require: $\Psi_i^j\ \forall i \mid 0 \leq j < n_i \mid 1 \leq n_i \leq 3$   ▷ Select $1-3$ limitations per aspect
Require: $O_{\Psi_i^j}\ \forall i,j;\ O_{\Psi_i^j} \in [0,10]$   ▷ Estimate how likely each limitation is to occur
Require: $S_{\Psi_i^j}\ \forall i,j;\ S_{\Psi_i^j} \in [0,10]$   ▷ Estimate how critical each limitation is
Require: $D_{\Psi_i^j}\ \forall i,j;\ D_{\Psi_i^j} \in [0,10]$   ▷ Estimate the likelihood of detection
$sum_\omega \leftarrow \sum_i \omega_i$
$\omega_i \leftarrow \frac{\omega_i}{sum_\omega}$
for each $i \in [0,5)$ do
    for each $j \in [0,n_i)$ do
        $T_i^j \leftarrow \sqrt[3]{O_{\Psi_i^j} \cdot S_{\Psi_i^j} \cdot D_{\Psi_i^j}}$
        if $O_{\Psi_i^j} = 10 \lor S_{\Psi_i^j} = 10 \lor D_{\Psi_i^j} = 10$ then
            $T_i \leftarrow 10$
        end if
        if $O_{\Psi_i^j} = 0 \lor S_{\Psi_i^j} = 0 \lor D_{\Psi_i^j} = 0$ then
            $T_i \leftarrow 0$
        end if
    end for
    $T_i \leftarrow \frac{1}{n_i} \sum_{j=0}^{n_i-1} T_i^j$
    for each $j \in [0,n_i)$ do
        if $T_i^j = 0$ then
            $T_i \leftarrow 0$
        end if
    end for
end for
$T \leftarrow \sum_{i=0}^{4} \omega_i \cdot T_i$
Ensure: $T \in [0,10]$   ▷ Resulting FRIES Trust Score $T$
```

The OSD scores are then multiplied, providing us with a Trust Score per aspect. The five resulting scores are weighed (by default they are equally weighed at 20%, but they can be case-specific) and the overall FRIES Trust Score is calculated this way. Compared to the standard FMEA, the evaluation scheme is inverted and ranges from 0 to 10. Thus instead of a high value being representative of high risks, a high score is representative of a highly trustworthy model or dataset. As the standard FMEA calculation favors small values (with random values for O, S and D, 92% of FMEA scores are ≤ 4.5), we need to remove this bias when inverting the score (45% of scores are ≤ 4.5), so that high values can be obtained, accordingly. To do this, we change the combination function from the usual $O \cdot S \cdot D$ to $\sqrt[3]{O \cdot S \cdot D}$. An example of what an evaluation with our adapted approach might look like can be seen in Tab. 8.

In addition, the option of assigning an optimal score of 10 has been added. An optimal score sets the resulting score for the current trust aspect to 10. This allows ignoring, for example, a risk that cannot occur (i.e., when O is set to 10, the probability of the occurrence is none, which means that the significance S of the non-occurring event is irrelevant). In the same vein, the possibility of assigning an absolute exclusion criterion with a score of 0 (what we call a deficit) has been added. Here, if even one of our aspects is evaluated with a critical 0, that means that the model or dataset as a whole should not be trusted. For a more formalized representation of the approach, we refer to Alg. 1.

3.2 Experimental Design

As a first attempt at testing the use of our metric, we have 10 machine learning researchers and users calculate the FRIES Trust Score of three models and three datasets each. Assuming that the definitions established in Section II are viable and the adapted FMEA approach previously described is applicable, only minor deviations between the user's rating of the models and datasets would be expected. The models and datasets are the following:

Datasets

- LARa (Logistic Activity Recognition Challenge) [35]: A motion capturing and IMU dataset of human activities performed in a warehousing scenario by 16 subjects. It provides activity classes of typical warehousing activities such as walking or handling goods. Beyond that, attributes such as gait cycle, left hand, right hand, or specific item poses have been annotated. They provide semantic descriptions of activities to facilitate transfer learning. The recording protocol is available online and the publication follows the FAIR-principle [49]. The identities of all subjects have been pseudonymized and anonymized. All subjects signed a consent form before recording.
- CelebA (Celebrity Attribute Dataset) [27]: A face attributes dataset containing images of celebrities. The images include annotations for various attributes such as hair color, age, gender, and facial expressions. CelebA is commonly used for tasks such as face recognition, attribute prediction, and facial attribute manipulation in computer vision research. It serves as a resource for training and evaluating machine learning models for face-related tasks and uses publicly available images of individuals who are prominent in the public eye.

- DukeMTMC (Duke University Multi-target Multi-camera) [40]: As previously mentioned, DukeMTMC was a widely used benchmark dataset for multi-target, multi-camera tracking challenges, such as the re-identification of pedestrians. The dataset contains annotated bounding boxes for pedestrians, along with their corresponding identities across different camera views. It was, however, retracted due to its data privacy violations.

Models

- GPT-3 (Generative Pre-trained Transformer) [5]: A natural language processing model, which belongs to the Transformer family of models. GPT-3 is trained on a vast amount of online text data and is capable of generating human-like text in response to prompts. It can perform a wide range of language tasks, including text completion, translation, question answering, and text or code generation. Neither the model itself, nor the data it is trained on are open source. The model is very well known, even beyond the research community and further sparked the conversation on trustworthy machine learning [41].
- YOLO (You Only Look Once) [39]: An object detection model, which is designed to detect objects within images or video frames by dividing the image into a grid and predicting bounding boxes and, in case of multiple models, class probabilities for each grid cell. Unlike its predecessors, YOLO does not require multiple passes through the network and instead processes the entire image in a single forward pass, making it faster and more efficient. Due to this, YOLO is still widely used and was adapted since its inception.
- GoogleNet [45]: Also known as Inception, GoogleNet is a widely used deep CNN developed by Google. It introduced the concept of the inception module for image classification, which incorporates multiple convolutions of different sizes and pooling operations within a single layer.

For all the above mentioned models and datasets the FRIES Trust Scores will be provided per user. The results obtained per aspect of each model and dataset will be represented by subscores as well. The participants remain anonymous. To rate the models and datasets, the participants were provided with a command line interface (CLI) script[4], that is hereby also provided to the research community.

4 Experimental Results

Having performed the experiments described in the preceding section, the scores attributed to the models and datasets by the participants for each respective trust aspect can be seen in Fig. 1. In some cases the scores per aspect vary greatly (e.g., DukeMTMC and GPT-3), while the subscores are fairly equally rated in other cases (e.g., YOLO and GoogleNet). There are apparent scoring

[4] www.github.com/KDD-OpenSource/trustscore

patterns, i.e., the participants rated the aspects of the model or dataset similarly – see LARa's fairness score or DukeMTMC's safety score. The only case in which the ten participant's ratings varied over nearly the entire scoring range is YOLO's safety score. For the dataset, a pattern is apparent as the explainability score is the highest-rated aspect for each dataset. The same pattern cannot be observed for the models tested.

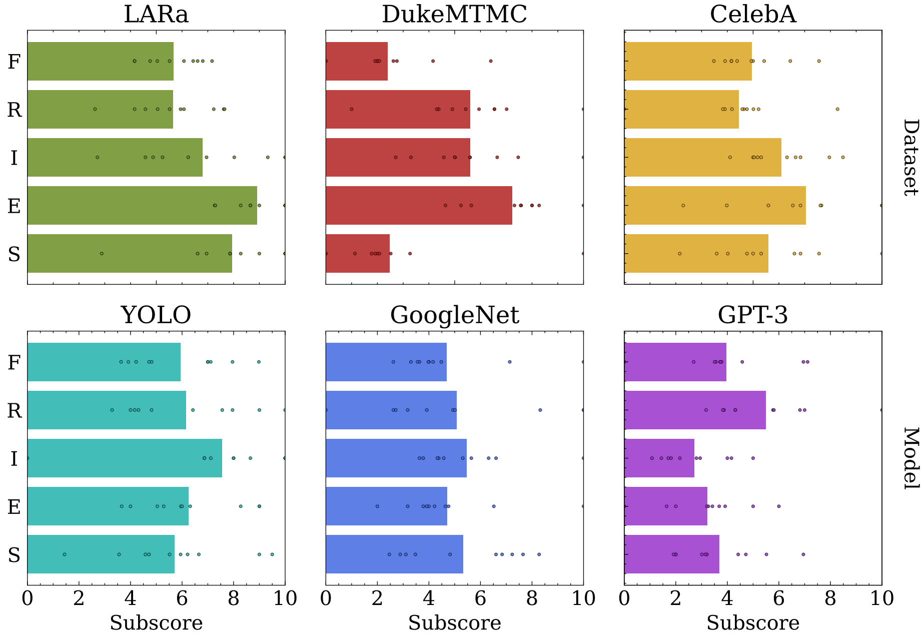

Fig. 1: The subscores per aspect, per model and dataset, as rated by the ten participants of our experiment.

The overall FRIES Trust Scores is depicted Fig. 2. The average scores are presented on the graph, the rectangles around it denote the 20% and 80% scoring quintile. As depicted in Fig. 1, the participants rated the models and datasets similarly in most cases, albeit, naturally with a certain spread. The spread is the smallest for GPT-3, potentially due to this model being very well known by all users. The highest spread is observed for DukeMTMC, with the 20% quintile starting at a FRIES Trust Score of 0. As described in the introduction, DukeMTMC is a dataset considered to be problematic, having been retracted due to privacy violations. Nevertheless, two of the participants scored the dataset with a FRIES Trust Score of 5. The dataset LARa was rated as the most trustworthy dataset with a FRIES Trust Score of 6.89. The highest rated model was YOLO, with a FRIES Trust Score of 5.81.

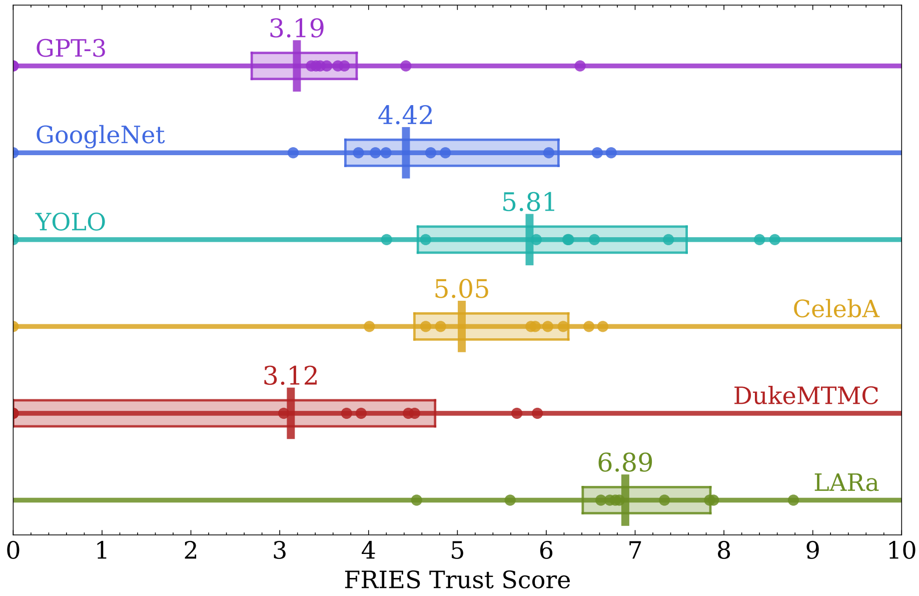

Fig. 2: The FRIES Trust Scores per model and dataset, as rated by the ten participants of our experiment.

Overall, the scores obtained through our experiment show apparent patterns and limited degrees of spread between the participants' ratings. During our experiments and the preceding research, it also became apparent, that some requisites of trustworthiness are contradictory. One such example is the added explainability of an accessible model, versus the added safety gained by securing a proprietary model.

Finally, the participants kindly provided us with feedback on the scoring procedure using our CLI script. The most common remarks were that the meaning of the OSD table could have been explained better, the risks could be explained more explicitly, and that an example of a rating of a model or dataset could have been provided for orientation. While scores of 0 and 10 were understood clearly, participants noted that especially the middle range of scores were ambiguous (i.e., that it is hard to gauge the difference between a significance of 4 and 5).

5 Conclusion and Outlook

In the beginning of this work, we set out asking whether DukeMTMC could be considered a trustworthy dataset, whether the criticism it encountered made it less trustworthy than others, assuming that data privacy and fair use represent an aspect of trustworthy machine learning.

To enable us and others to answer such questions about DukeMTMC and other datasets and models, we performed a literature review and, based on the results of this research, determined the most integral aspects of trustworthy machine learning. Subsequently, we propose first operational definitions of these aspects, namely fairness, robustness, integrity, explainability, and safety.

To quantify these aspects of trustworthiness, we then developed an approach, inspired by FMEA, permitting us to rate models and datasets with a trust score that we call the FRIES Trust Score. 10 users applied this approach to three models and datasets, rating them on a 0 to 10 FRIES Trust Scale. It could be observed that, even though outliers are present, pronounced overlaps between the participants ratings of the models and datasets could be observed.

Concerning contradictory requirements of the subscores, the question remains, whether the former can be reconciled. In addition, the results obtained are still limited in their meaningfulness, as our experiments were only small-scale. A certain spread between the results can still be observed, which should be further minimized as well.

Participants provided feedback on the clarity of the rating procedure, asking for more explicit descriptions of risks and an example of a rating, for orientation purposes. As a next step, after having performed a deductive category formation and a first analysis of these categories through our experiments, an inductive category formation could be performed.

The FRIES Trust Score is the first measure to quantify trust in machine learning. Based on our findings, we believe the FRIES Trust Score to be well suited to be first stride towards a unified and operationalized trust score for machine learning models and datasets. Thus, the research community is invited to reproduce our results and to apply our approach to state-of-the-art as well as their own models and datasets. For this purpose, we provide a CLI script, with which interested readers can reproduce our experiments themselves. We also greatly support [11]'s notion that assuring fairness (and to our understanding trust overall) remains a human-in-the-loop process. Thus, in our opinion, subjectivity will remain a part of the operationalization process of trust for the foreseeable future.

References

1. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
2. Ashoori, M., Weisz, J.D.: In AI we trust? factors that influence trustworthiness of AI-infused decision-making processes. arXiv preprint arXiv:1912.02675 (2019)
3. Ashton, R.H.: The robustness of linear models for decision-making. Omega **4**(5), 609–615 (1976)
4. Auer, P., Holte, R.C., Maass, W.: Theory and applications of agnostic pac-learning with small decision trees. In: Machine Learning Proceedings 1995, pp. 21–29 (1995)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)

6. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)
7. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Computing Surveys (2020)
8. Dalenius, T.: Towards a methodology for statistical disclosure control (1977), publisher: Statistics Sweden
9. Darestani, M.Z., Chaudhari, A.S., Heckel, R.: Measuring robustness in deep learning based compressive sensing. In: International Conference on Machine Learning. pp. 2433–2444. PMLR (2021)
10. De Cristofaro, E.: An overview of privacy in machine learning. arXiv preprint arXiv:2005.08679 (2020)
11. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Dugan, C.: Explaining models: An empirical study of how explanations impact fairness judgment. In: 24th International Conference on Intelligent User Interfaces. pp. 275–285 (2019)
12. Drenkow, N., Sani, N., Shpitser, I., Unberath, M.: A systematic review of robustness in deep learning for computer vision: Mind the gap? arXiv preprint arXiv:2112.00639 (2021)
13. European Commission: 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), https://eur-lex.europa.eu/eli/reg/2016/679/oj
14. European Commission: Amendments adopted by the european parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) (2023), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:C_202400506
15. Goodfellow, I., McDaniel, P., Papernot, N.: Making machine learning robust against adversarial inputs. Communications of the ACM **61**(7), 56–66 (2018)
16. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys **51**(5), 1–42 (2018)
17. Haber, S., Stornetta, W.S.: How to time-stamp a digital document. Journal of Cryptology **3**(2), 99–111 (1991)
18. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems **29** (2016)
19. He, Z., Zhang, T., Lee, R.B.: Verideep: Verifying integrity of deep neural networks through sensitive-sample fingerprinting. arXiv preprint arXiv:1808.03277 (2018)
20. Herm, L.V., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. International Journal of Information Management **69** (2023)
21. Hou, R., Ai, S., Chen, Q., Yan, H., Huang, T., Chen, K.: Similarity-based integrity protection for deep learning systems. Information Sciences **601**, 255–267 (2022)
22. Ji, Z., Lipton, Z.C., Elkan, C.: Differential privacy and machine learning: a survey and review. arXiv preprint arXiv:1412.7584 (2014)
23. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in Neural Information Processing Systems **30** (2017)
24. Kuttichira, D.P., Gupta, S., Nguyen, D., Rana, S., Venkatesh, S.: Verification of integrity of deployed deep learning models using bayesian optimization. Knowledge-based Systems **241** (2022)

25. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: From principles to practices. ACM Computing Surveys (2023)
26. Liu, J., Wang, D., Lin, Q., Deng, M.: Risk assessment based on fmea combining dea and cloud model: A case application in robot-assisted rehabilitation. Expert Systems with Applications **214** (2023)
27. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE international Conference on Computer Vision (ICCV). pp. 3730–3738 (2015)
28. M. Strobel, R. Shokri: Data privacy and trustworthy machine learning. IEEE Security & Privacy **20**(5), 44–49 (2022)
29. Marcinkevičs, R., Vogt, J.E.: Interpretability and explainability: A machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805 (2020)
30. Mayring, P.: Qualitative content analysis. A companion to qualitative research **1**(2), 159–176 (2004)
31. Meske, C., Bunde, E.: Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In: Artificial Intelligence in HCI. pp. 54–69 (2020)
32. Mitchell, T.M.: The need for biases in learning generalizations. Rutgers University (1980)
33. Mohseni, S., Wang, H., Xiao, C., Yu, Z., Wang, Z., Yadawa, J.: Taxonomy of machine learning safety: A survey and primer. ACM Computing Surveys (2022)
34. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
35. Niemann, F., Reining, C., Moya Rueda, F., Nair, N.R., Steffens, J.A., Fink, G.A., ten Hompel, M.: LARa: Creating a dataset for human activity recognition in logistics using semantic attributes. MDPI Sensors **20**(15) (2020)
36. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P.: Sok: Security and privacy in machine learning. In: IEEE European Symposium on Security and Privacy (EuroS&P). pp. 399–414 (2018)
37. Peng, K., Mathur, A., Narayanan, A.: Mitigating dataset harms requires stewardship: Lessons from 1000 papers. NeurIPS 2021 Datasets and Benchmarks Track (2021)
38. Rauber, J., Brendel, W., Bethge, M.: Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:1707.04131 (2017)
39. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
40. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision (2016)
41. Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., Pauly, M.: The self-perception and political biases of ChatGPT. Human Behavior and Emerging Technologies (2024)
42. Schmitz, A., Akila, M., Hecker, D., Poretschkin, M., Wrobel, S.: An approach for systematic quality assurance when working with ML components. AT - Automatisierungstechnik **70**(9), 793–804 (2022)
43. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: 30th USENIX Security Symposium. pp. 2615–2632 (2021)
44. Stamatis, D.H.: Risk management using failure mode and effect analysis (FMEA). Quality Press (2019)

45. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015)
46. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electronic Markets **31**, 447–464 (2021)
47. Thuraisingham, B.: Trustworthy machine learning. IEEE Intelligent Systems **37**(1), 21–24 (2022)
48. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., van Moorsel, A.: The relationship between trust in AI and trustworthy machine learning technologies. In: ACM Conference on Fairness, Accountability, and Transparency (FaccT). pp. 272–283 (2020)
49. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**(1), 1–9 (2016)
50. Wing, J.M.: Trustworthy AI. Communications of the ACM **64**(10), 64–71 (2021)
51. Wischnewski, M., Krämer, N., Müller, E.: Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (2023)
52. Xu, J., Chen, J., You, S., Xiao, Z., Yang, Y., Lu, J.: Robustness of deep learning models on graphs: A survey. AI Open **2**, 69–78 (2021)
53. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: 30th International Conference on Machine Learning (ICML). vol. 28, pp. 325–333 (2013)