Methods

Anna Schmitz*, Maram Akila, Dirk Hecker, Maximilian Poretschkin and Stefan Wrobel

The why and how of trustworthy AI

Vertrauenswürdige KI – Warum und Wie?

An approach for systematic quality assurance when working with ML components

Ein Ansatz zur systematischen Qualitätssicherung beim Einsatz von ML Komponenten

https://doi.org/10.1515/auto-2022-0012
Received February 4, 2022; accepted April 25, 2022

Abstract: Artificial intelligence is increasingly penetrating industrial applications as well as areas that affect our daily lives. As a consequence, there is a need for criteria to validate whether the quality of AI applications is sufficient for their intended use. Both in the academic community and societal debate, an agreement has emerged under the term "trustworthiness" as the set of essential quality requirements that should be placed on an AI application. At the same time, the question of how these quality requirements can be operationalized is to a large extent still open.

In this paper, we consider trustworthy AI from two perspectives: the product and organizational perspective. For the former, we present an AI-specific risk analysis and outline how verifiable arguments for the trustworthiness of an AI application can be developed. For the second perspective, we explore how an AI management system can be employed to assure the trustworthiness of an organization with respect to its handling of AI. Finally, we argue that in order to achieve AI trustworthiness, coordinated measures from both product and organizational perspectives are required.

***Corresponding author: Anna Schmitz,** Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany, e-mail: anna.schmitz@iais.fraunhofer.de, ORCID: https://orcid.org/0000-0001-8801-3700
Maram Akila, Dirk Hecker, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany; and ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany, e-mails: maram.akila@iais.fraunhofer.de, dirk.hecker@iais.fraunhofer.de
Maximilian Poretschkin, Stefan Wrobel, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany; and ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany; and Institut für Informatik III, Rheinische-Friedrich-Wilhelms-Universität Bonn, Bonn, Germany, e-mails: maximilian.poretschkin@iais.fraunhofer.de, stefan.wrobel@iais.fraunhofer.de

Keywords: trustworthy AI, AI quality assurance, AI management systems, AI testing and validation

Zusammenfassung: Künstliche Intelligenz findet zunehmend in der Industrie sowie in Bereichen unseres Alltagslebens Anwendung. Daher werden Qualitätsstandards benötigt, um eine KI-Anwendung in ihrem Einsatzkontext zu bewerten. In der Forschung und gesellschaftlichen Debatten zeichnet sich eine Übereinkunft zu grundlegenden Qualitätsanforderungen an KI-Anwendungen ab, welche meist unter dem Begriff der „Vertrauenswürdigkeit" zusammengefasst werden. Die Operationalisierung dieser Qualitätsanforderungen wiederum ist noch in weiten Teilen offen.

Im vorliegenden Beitrag betrachten wir vertrauenswürdige KI aus zwei Perspektiven: Produkt- und Organisationsperspektive. Für erstere stellen wir einen Ansatz zur KI-spezifischen Risikoanalyse vor und skizzieren, wie eine Argumentationskette für die Vertrauenswürdigkeit einer KI-Anwendung entwickelt werden kann. Aus der Organisationsperspektive betrachten wir, wie ein KI Managementsystem zur Sicherstellung eines vertrauenswürdigen Einsatzes von KI in einer Organisation beitragen kann. Schließlich zeigen wir auf, dass die Umsetzung vertrauenswürdiger KI abgestimmte Maßnahmen aus beiden Perspektiven erfordert.

Schlagwörter: Vertrauenswürdige KI, KI Qualitätssicherung, KI Management Systeme, KI Testen und Validierung

1 Introduction

Artificial intelligence (AI) is penetrating numerous areas of our lives at a breathtaking speed. In the process, it is increasingly taking over central activities, on the one hand in safety and security-relevant domains, such as automated driving [60], but also in domains that affect human self-determination, such as in the granting of loans [20] or in hiring processes [2]. It is generally assumed that AI will

only deploy its full potential if it is used in accordance with high quality standards and in line with our core values. In this context, the concept of trustworthiness has been coined, which has already been the subject of intensive research for several years [16], [17], [54]. For European practice, the Ethics Guidelines for Trustworthy AI of the High-Level Exert Group [23] published in 2019 are directional and it can be expected that their specifications will find their way into legal requirements through the recently published proposal for AI regulation of the European Commission [15] (for an analysis of the proposal, see [51]). It is noteworthy to mention that the High-Level Expert Group's guidelines, similarly as numerous other AI ethics statements [9], leave open which entity has to comply with their requirements and how. Comparisons to related but simpler domains, such as IT security, make it seem plausible that trustworthiness can and must refer to the organizations that use or develop AI as well as to the AI systems themselves. Looking at the concrete example of "diversity, non-discrimination and fairness," the HLEG recommends establishing a mechanism that allows others to flag issues related to bias and discrimination [23], which is a non-technical measure of the organization, while the recommendation to monitor fairness through metrics [23] can only be addressed by a technical examination of the specific AI application. For both interpretations, the operationalization of trustworthiness requirements is to a large extent still open [21], [43]. This paper takes both perspectives with a focus on AI applications which are based on Machine Learning (ML) technologies, presents approaches for operationalization, and discusses how they are related.

2 The why of trustworthy AI

The impressive progress of AI in recent years, for instance in medical diagnostics [59] and predictive maintenance [8], was largely driven by Machine Learning. The great potential of ML-based applications originates from the fact that their behavior is not predefined by programmed rules but inferred from data instead. More specifically, ML methods are designed to identify patterns in so-called training data and incorporate their statistical insights in a model. Deep neural networks in particular, which are a type of complex ML models generated from vast amounts of data, are capable of performing tasks that, up to date, could not be described effectively by human rules (e. g., image [12] and speech recognition [4]). However, the data-driven nature of ML gives rise to novel imponderables. For instance, ML-based applications bear the risk to incorporate and reproduce discriminatory patterns in real data [42], by which individuals as well as the reputation of the organization in charge might adversely be affected. This example illustrates why the recommendations and requirements in the numerous AI guidelines presented (see chapter 1) are not solely aimed at preventing bodily injury, property damage or financial loss, as it is the case with "classic" IT security and safety. Complementary, parts of them also consider the (intangible) impact of AI on individuals and society, given, for example, its increasing use to support sensitive decision-making processes (e. g., dismissal of workers [9] or sentencing delinquent persons [1]).

Given the over 700 AI policy initiatives from 60 countries and territories, as well as over 170 emerging AI-related regulation initiatives [47], it is apparent that there are different views around the world on how AI should be used. This partially originates from the diversity of national and territorial AI strategies driving official guidelines and AI regulation e. g., [14], [55], [46], and [52]. While requirements are globally different, one finds recurring elements in numerous AI guidelines, especially when looking at western initiatives. More precisely, there is a consensus among the variety of European as well as US American guidelines that privacy, fairness and transparency (see [34] and [64]) are also, in addition to the established objectives of reliability, safety, and security, essential quality dimensions of AI. Moreover, the EU level documents at which we focus in the following emphasize human autonomy and oversight (partially in contrast to other international perspectives) as a further key requirement for trustworthy AI. While reliability, safety, and security address the proper functioning and potential vulnerabilities of the system (see #2 [23] and Art. 15 [15]), especially the data-driven nature of ML, as mentioned, gives rise to novel aspects. Clearly, as soon as personal data is processed, the privacy of the individual needs to be protected. Here, the use of ML involves risks such as personal data being newly generated during operation (e. g., personal keystroke patterns [37]) or extracted from the ML model without authorization [48], [18]. This notion can be extended to other forms of data e. g., business sensitive data. Furthermore, fairness must be ensured in decision processes (see #5 [23] and Art. 10, 15 [15]), considering that ML algorithms might learn from a biased ground-truth or have a significantly lower performance with respect to certain groups of people [6] (e. g., due to inconsistent data quality). Following, as stated, data-driven AI thrives on concepts that are hard to understand or ill-defined for humans. It is therefore not surprising that often the presented solutions,

i. e., ML models, are equally incomprehensible even for experts [7]. For cases were sound argumentations on quality or specific properties are needed, additional goals towards explainability can thus be desirable (see #4 [23] and Art. 13 [15]). As the terminology is not always consistent among the different documents, nor are there precise definitions in all cases, we consider (technical) explainability one aspect of transparency in the context of AI, while transparency is related to further aspects such as traceability (see #4, 7 [23] and Art. 12 [15]) and the provision of information (see #1, 2, 4, 7 [23] and Art. 13, 52 [15]). Finally, the High-level Expert Group and the European Commission have emphasized human autonomy and oversight among the most important requirements for the use of AI. While the former addresses risks stemming from the sheer fact that the automation provided by an AI application may (in some form) limit or manipulate the agency of humans (see "fundamentals rights impact analysis" [23] and Art. 5, 6 [15]), the latter aims at ensuring effective human supervision and, if necessary, intervention (see #1 [23] and Art. 14 [15]). As such, it appears that a broader conception of quality has been adopted in the context of AI, which is well captured by the term "trustworthiness".

While a consensus is emerging on the key dimensions of trustworthy AI especially at European level, for a given AI application, practical test procedures are needed to evaluate whether the corresponding quality characteristics are actually present. Notably, the research regarding the measurement and implementation of AI trustworthiness is broad and ongoing [24], [22], [45]. However, while it is already apparent that effective measures to ensure properties such as reliability, fairness, or transparency often require access to the training data, the design process, and the representation of the output, there are still several challenges to be addressed in the concrete verification of requirements [62], [5]. For example, as individual parts of a complex ML model are typically challenging to interpret [63], quality aspects cannot be broken down (see [39] for a discussion on single component evaluations). Further challenges arise as AI applications often operate in complex environments which cannot be fully understood and captured by humans themselves (e. g., the indicators of diseases in medical data). Thus, often concise technical criteria are missing to evaluate the coverage of training and test data (see [19] for a combinatorial perspective). In open-world contexts in particular, it is typically not possible to quantify the application domain precisely. Looking further at the example of reliability, as opposed to other engineering disciplines such as construction where one can guarantee by computational proof (or reach a very high assurance level at least) that a building is statically safe – provided no huge earthquake occurs –, the verification of ML-based applications proves challenging. While promising approaches to attest the flawless functionality of an ML model are being developed (e. g., SMT-solver [35], linear relaxation [13], and branch-and-bound algorithms [61], see also [40] for a survey on verification algorithms), they are not yet sufficiently advanced to be of use for practical AI applications [5]. Additional complexity emerges in case the ML models are continuously trained during operation. Notably, even if the challenge of conducting an AI system audit was solved, still, giving a warranty appears difficult given the dynamics of ML. At the same time, however, AI systems are currently being applied to many areas of our lives whose trustworthiness must be ensured.

3 The how of trustworthy AI

As elaborated in the previous chapter, trustworthiness has been adopted as a broad notion of quality for AI applications, exceeding the classic concepts of IT security and safety. While requirements and recommendations that characterize "trustworthy AI" have been presented in numerous guidelines [34], they are, however, introduced at an abstract level and leave to a large extent open which measures or methods should best be taken to meet them [21], [43]. Overall, there is a lack of standards that would concretize requirements [11], and technical verification remains a challenge (see chapter 2). Parallelly, numerous organizations already use AI in practice and have high quality expectations for their systems. Thus, it is crucial that they take into account the imponderables caused by AI and find a systematic way of dealing with them. To this end, we observe that two fundamental perspectives must work together: i) a high technical quality of AI systems themselves is required, ii) the organization should make appropriate preparations (e. g., establish structures, processes and roles) for handling its AI applications and their development in a trustworthy manner. In the following, we refer to these two perspectives as the product and organizational perspectives.

The product perspective, detailed in Section 3.1, focuses on an individual "product", meaning here an AI system, within its application context. The approach presented, serves to systematically evaluate the trustworthiness of an AI application with respect to the dimensions mentioned in the previous chapter. It is based on a structured analysis of risks and, moreover, involves their mitigation through mostly technical but, in parts, also non-technical measures along the lifecycle of the application.

Notably, risks can extend well beyond the initial rollout, either because they were missed during conception of the system or did not even exist at the time. Such continuous or latent issues need to be addressed at the organizational level and the corresponding (organizational) superstructure is discussed in Section 3.2 (we refer to [53] and [41] for the importance of management support for achieving organizational objectives in the implementation of classic IT systems). This superstructure not only needs to reflect long-lasting risk situations, but in addition governs the general policies of the organization towards the dimensions of trustworthy AI and orchestrates general, often process-oriented, measures to maintain these policies in practice. We introduce AI management systems as potential candidates for such a superstructure and explore their role in operationalizing trustworthy AI from the organizational perspective.

3.1 Approach from the product perspective

Judging by the notion of trustworthiness endorsed by current EU-level documents, it is apparent that trustworthy AI can, in general, only be achieved by a broad analysis of both the application and its environment, accompanied by careful risk mitigation measures and checkups. However, existing test procedures for traditional IT systems cannot be straightforwardly used for AI (see chapter 2) and, moreover, do not address intangible risks related to discrimination and human autonomy, among others. For these reasons, novel technical foundations are needed for the assessment and control of AI risks. Below, we elaborate on the challenges in this regard and present an approach for the systematic evaluation of trustworthy AI, putting a particular focus on AI applications based on ML technologies.

Since the recommendations and requirements for trustworthy AI are abstract [21], [43], they must first be further operationalized in order to evaluate a specific AI application. Here, we observe three main challenges: First, various aspects of trustworthiness are not readily quantifiable by metrics or key performance indicators. For instance, it is unclear how to measure the user autonomy afforded by an AI application or the degree of its explainability. Second, typically, the concrete requirements and evaluation standards strongly depend on the application context as well as the specific AI method used. For example, while fairness is irrelevant for use cases in production plants, there are high risks in applications that support critical decisions such as algorithms for credit scoring or rating of job applicants. Following on this example, over 20 fairness metrics have been proposed [56], each of them quantifying a different aspect of discrimination. However, there is no statement by official bodies about which metric(s) should be applied for a given use case and, furthermore, which target interval to rate as acceptable. Third and lastly, disparate objectives associated with trustworthiness may conflict with one another leading to conflicting system-level requirements. One reason is that numerous requirements which reflect societal values are detached from performance-related quality requirements. For instance, a perfect predictor which violates the principle of fairness given a discrimination-inherent ground truth, creates a conflict between non-discrimination and optimizing its performance. Similarly, the prediction capabilities of an ML model may be compromised if personal information is removed from the training dataset for privacy reasons.

Interestingly, some of the aforementioned challenges can also be found, for example, in IT Security and Functional Safety, where the objectives of resistance to manipulation or malfunction often lead to very different quality criteria for different systems [58], [3]. Similar to the risk-based approaches established in these fields, we believe that the specification of trustworthiness criteria for an AI application must be based on a comprehensive risk analysis. Moreover, given the variety of methods for risk mitigation, the measures for meeting defined criteria must also be adapted to its respective AI application. Accordingly, for assuring and evaluating a trustworthy AI, we propose an approach consisting of two consecutive phases: i) top-down and ii) bottom-up. While the top-down phase serves to identify and break down risks into specific criteria for the AI system, the bottom-up phase encompasses the (likewise specific) measures taken that work towards meeting these criteria and argues for their fulfillment. In the following, we will briefly present the main steps of this approach which we have described in more detail in [49] and [50].

As a basis for systematically analyzing and breaking down AI risks in the top-down phase, especially those stemming from the use of ML, we propose a scheme along the six dimensions presented in chapter 2 (see Figure 1). Since each dimension covers numerous aspects which must be weighted and evaluated differently depending on the application context, it is necessary to subdivide them respectively into more granular risk areas [49]. For instance, regarding privacy, different evaluation standards with respect to personal and business sensitive data may emerge. Similarly, when defining specific criteria with respect to transparency, a distinction should be made between explainability for technical experts and information of users and affected persons, for example. Finally, our proposed scheme includes dedicated risk areas “control of

Dimension	Risk area
Fairness	Fairness
	Control of dynamics
Autonomy and Control	Distribution of tasks between human and AI system
	Information and empowerment of users and stakeholders
Reliability	Regular operation
	Robustness
	Evasion strategies
	Uncertainty
	Control of dynamics

Dimension	Risk area
Privacy	Protection of personal data
	Protection of business sensitive information
	Control of dynamics
Transparency	Transparency for users
	Explainability for experts
	Auditability
	Control of dynamics
Safety and Security	Functional safety
	Integrity and availability
	Control of dynamics

Figure 1: AI risk scheme. For a detailed elaboration on the dimensions, see [10], and on the risk areas, see [49].

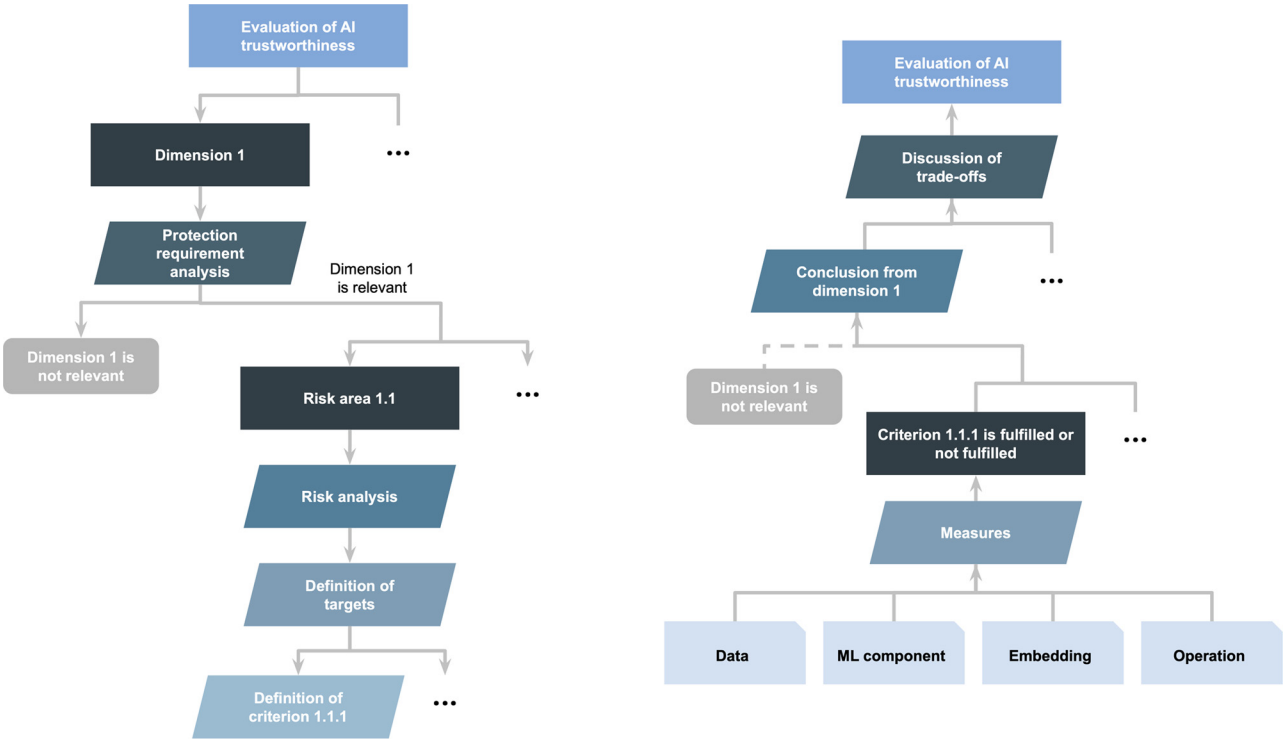

Figure 2: Visualization of the product perspective approach. The l.h.s shows the top-down phase, the r.h.s the bottom-up phase.

dynamics" to take into account risks which arise during operation such as model and data drift.

The top-down phase of our approach (see l. h. s. Figure 2) serves to specify trustworthiness criteria for measuring the control of the AI risks in the scheme presented. To increase efficiency, this phase begins with a protection requirement analysis at the dimension level, where an evaluation of the potential damage determines whether a dimension is relevant for the considered AI application. Exclusively for relevant dimensions, an in-depth analysis for each associated risk area follows. Based on the results, targets are formulated describing under which circumstances residual risks are acceptable. These targets are then translated into (preferably quantitative) criteria for the specific

AI application. Let us illustrate on the example of ML-based credit scoring. Since this application has an impact on the financial capabilities of individuals, most likely fairness will be among the relevant dimensions. The in-depth risk analysis should be informed e. g., by the applicable anti-discrimination legislation, and help to identify input features which characterize potentially disadvantaged population groups. Lastly, use case-adequate fairness metrics (and target values, if applicable) are chosen, such as statistical parity and counterfactual fairness [57].

After defining the trustworthiness criteria for the AI application, the bottom-up phase of our approach assures and argues that they are fulfilled. To this end, measures (preferably technical) for risk mitigation and testing, which often need to be specifically adapted to the AI technology used, should be implemented, where applicable. We distinguish between four categories of measures, see Figure 2 (r. h. s.): i) data, ii) ML component, iii) embedding, and iv) operation. These categories, however, should not be considered separately as they represent different starting points for risk control along the life cycle. Since our approach focusses on AI applications whose functionality is typically derived directly from data, the selection and quality of data are of particular importance. Complementary, risks should be considered during the design and training procedure of the ML component included in the system (meaning the ML model and, if applicable, pre- and post-processing). Moreover, in numerous applications, the ML component operates in interaction with further, non-ML components (e. g., rule-based), which may serve for logging, monitoring or otherwise processing in- or outputs. Unlike for the ML component, conventional measures and tests can be used for this embedding. Lastly, ensuring the trustworthiness of an AI application does not end at the development stage since various levels of support and maintenance are needed during operation. This cannot, or rather should not, be addressed by technical solutions alone, but involve human oversight as well as allow for the option of human intervention in critical situations. Returning to the example of credit scoring, several measures can be employed to work towards fulfilling set target values with the fairness metrics chosen, for example bias-removing pre-processing of the training data or post-processing of the ML model outputs. Moreover, a monitoring of fairness indicators should be installed to check e. g., for potential deviations from the target values during operation.

The bottom-up phase (see r. h. s. Figure 2) encompasses the implementation of (technical as well as non-technical) measures with respect to the four abovementioned categories. Based on their documentation, a safeguarding argumentation is developed for each risk area under consideration in order to demonstrate, if possible, that the criteria from the top-down phase are met. Subsequently, a conclusion is drawn per dimension on whether the criteria are met or not. Finally, if residual risks still exist due to conflicting requirements between the dimensions for example, trade-offs should be discussed and carefully weighed prior to the final judgement about the application's trustworthiness.

3.2 Approach from the organizational perspective

Looking at the recommendations and requirements for trustworthy AI in detail, one quickly recognizes that meeting them is not limited to the AI system, but also involves the organization in charge, even if it is not explicitly demanded. This, for example, becomes apparent in the discussion of trade-offs as mentioned in the previous section: While the product perspective provides technical foundations for analyzing and treating risks in relation to a specific AI application, it cannot generally determine how conflicts between system requirements should be resolved. For instance, in the example of credit scoring above a trade-off between performance (reliability) and fairness might exist. Decisions on such trade-offs should be guided by organizational policies and, where applicable, take into account further requirements such as non-AI sector regulation or budget. In general, this also affects which AI risks the organization considers acceptable. Parallelly, setting up standardized processes and procedures also for the implementation level, can help to effectively achieve a consistent level of trustworthiness. Below, we will present how a trustworthy handling of AI can be approached in a structured manner and, if successful, demonstrated using a dedicated AI management system.

A generally established means for organizations to achieve defined objectives purposefully and accountably, are management systems (MS'). Following [33], MS' constitute the "set of interrelated or interacting elements of an organization to establish policies and objectives, as well as processes to achieve those objectives" [ibid.]. Here, it is initially not specified which goals are to be achieved; for example, management systems are being used for assuring information security [27], product or service quality [28], and environmental sustainability [29] within organizations. Aiming at combining all organizational units and processes which are directed to set objectives into a clear framework, a MS inevitably affects the entire organization, from leadership to specific technical and organizational

measures. Typical parts or tasks within a management system (as stated in [36], [44]) include:

- formulating objectives in the form of policies,
- analyzing risks and opportunities for these objectives,
- defining roles or responsibilities for specific (sub-)objectives,
- (...) planning and implementing processes and the measures required to achieve them,
- and planning, implementing and evaluating reviews of the achievement of objectives.

Given the above description, an appropriately designed management system could be a suitable manner to create an environment where it is ensured that the technical as well as non-technical requirements for trustworthy AI are met. Unlike in the product perspective, such a MS is at first detached from individual applications but can eventually impact their trustworthiness. Also the European Commission appears to consider MS' as an appropriate approach, since it requires both a risk management system and a quality management system for providers of high-risk AI systems (Art. 9 and 17 [15]). While the proposal for AI regulation lists basic processes and activities that such MS' should encompass, their specific design, however, needs to be further operationalized. To provide effective support for an organization in handling their AI applications in a trustworthy way, we believe that a management system should be AI specific in each of the points listed above. First, starting from the leadership level, the trustworthiness dimensions as discussed in chapter 2 should be reflected in policies and made aware within the organization. Second, the notion of (organizational) risk must be extended to the broad concept of trustworthy AI. For example, a credit scoring application systematically refusing loans for women while approving for men with equal financial provision can cause financial damage to the provider and, foremost, harms the personal rights of women. Here, by formulating guidelines for risk analysis (e. g., based on the approach presented in Section 3.1), an AI management system can help not only to consider potential (negative) effects on the organization itself but also to examine how its AI products might affect individuals and society. In addition, as mentioned, decisions must be taken (informed by organizational policies) regarding risk acceptance levels and specific trustworthiness criteria. To this end, procedures such as stakeholder consultation for choosing use case-adequate fairness metrics might be needed. Third, being accountable for decisions is an essential aspect of a trustworthy AI handling. Given the residual risks which might exist in the use of AI, responsibilities and liabilities must be assigned to certain roles within the organization. Moreover, regarding the challenge that AI trustworthiness combines diverse objectives which may conflict with each other (e. g., performance vs. fairness, see Section 3.1), a committee should be established such as an "Ethical AI Review Board" to discuss, agree upon, and be accountable for trade-offs [23]. Fourth, as indicated above, numerous technical foundations emerging from the product perspective require an organizational counterpart i. e., an activity or process (preferably standardized) that puts them into practice. Notably, due to the dynamics inherent to ML, "post-market monitoring" is crucial (Art. 61 [15]). Furthermore, organizations must deal with the fact that AI specific risks cannot usually be completely safeguarded, and unforeseen errors may occur. In this respect, also non-technical measures might be needed such as mechanisms that allow to flag issues, for example related to discrimination, or allow for redress [23]. In order to create an AI-sensitized environment, these (non-)technical implementation levels should be accompanied by adequate resource planning (e. g., to assemble diverse developer teams) as well as training measures. Lastly, specific risks and challenges, such as fairness or data quality in general, ask for continuous and critical reevaluation. Audits of an AI application (internal or external) must be carefully planned and prepared by the organizations which, for instance, needs to formulate technical documentations (particularly if the proposal for AI regulation is adopted, see Art. 11 [15]).

The success of management systems in other domains (such as IT security, product quality, see [32]), in parts, can be attributed to their standardization, which captures best practices and thus enhances operationalization as well as comparability across organizations. For instance, the international standard for quality management systems (QMS) is being used for over 30 years and more than 1 million certificates have been issued [38]. Interestingly, given that the European Commission requires management systems for both quality and risk (see above), risk management has been standardized as well [31]. However, neither of the standards mentioned provides guidance for organizations on how to deal with the specific challenges and novel risks that arise from the use of AI. Therefore, a joint working group (ISO/IEC JTC 1/SC 42/WG 1) of the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC) is developing a set of standards to address different aspects of the use of AI technologies. Two particular ones that are currently under development are the international standards for AI risk management [25] and AI management systems [26].

Judging by the Working Draft of the international standard for AI MS, it proposes a suitable tool for organizations

to address the requirements for trustworthy AI presented in [23] and [15] (for an in-depth analysis, we refer to [44]), despite not formally being a QMS as demanded in [15]. While the main part of the standard defines requirements for the basic framework of an AI MS, the annexes provide controls which are recommendations for its implementation (whereby it should be noted that technical aspects are only treated by controls). For example, the working draft lists "fairness, security, safety, privacy, transparency and explainability, accountability, availability, maintainability, quality of training data, and AI expertise" as possible AI-related organizational objectives when managing risks (Annex B, [25]), thus providing a direction for risk management towards considering the dimensions of chapter 2. Another recommendation in this standard is to integrate an impact assessment into the organization's risk analysis process. Interestingly, the document introduces the term "impact" explicitly in the context of health and safety of society, as well as traditions, values, the environment, privacy, and fairness, among other things, thus extending the concept of organizational risk to the notion of trustworthiness. Notably, the request for impact assessment differentiates this Working Draft from other MS standards, making clear that such standards would not be sufficient to ensure a trustworthy handling of AI within an organization.

Lastly, it is noteworthy that MS standards do not only provide support for organizations to achieve their goals in a responsible and accountable manner, but they are also a generally accepted means to generate evidence for this. Accordingly, based on the previous discussion, we consider certification of an AI management system to be the basis for trustworthy AI and a suitable way for organizations to demonstrate their trustworthiness. Nonetheless, compliance with a MS standard is not sufficient for assuring the trustworthiness of an AI system itself. As technical verification cannot be waived, there are still numerous challenges to be overcome in order to conduct a comprehensive audit on the system level.

4 Conclusion

In this paper, we have outlined essential quality dimensions of trustworthy AI and argued that two different perspectives are involved in their operationalization, namely the product and organizational perspective. While the product perspective deals with the risks emerging from a specific AI application, the organizational perspective considers the handling of AI systems in view of organizational objectives and risks. For each of the perspectives, we have presented an approach for operationalizing and assuring trustworthy AI. The first approach serves to systematically evaluate an AI application from a product perspective. The second approach presents a management system specific to AI. Overall, we conclude that AI trustworthiness can only be achieved through a harmonized interplay between product and organizational perspectives since risks emerging from the product must be reflected in organizational policies as well as monitored, and in parts also mitigated by corresponding processes. In addition, as shown in both approaches presented, we believe that the implementation of concrete measures, from either perspective, should be based on a comprehensive risk analysis and specifically adapted to the respective results.

Our discussion of trustworthy AI focuses on ML-specific key dimensions for which there is a consensus in both academia and society. Furthermore, there are other important principles such as sustainability and environmental friendliness ("underrepresented compared to other ethical dimensions" [34]) which could be harmed in the use of ML, for example, due to the high energy consumption in training models. We believe that these principles are generally related to numerous non-ML-specific aspects and, thus, cannot be solely addressed from a product perspective, but other approaches are needed. For instance, environmental sustainability cannot be assessed without consideration of the resources from which electricity is generated.

Regarding the two perspectives introduced in this paper, we observe that the operationalization from the organizational perspective is further advanced since best practices are gradually emerging with respect to processes in AI development and operation (e. g., facilitating automation and tracking via MLflow, DVC, Jenkins) as well as AI management (as reflected in current standardization activities [25], [26]). In a next step, audit and certification schemes are needed, and it can be assumed that, once processes and procedures have been standardized, these schemes can be developed and applied in a similar fashion as for non-AI domains. From a product perspective, numerous testing tools are available for assessing specific trustworthiness-related characteristics in AI systems despite the challenges presented (see chapter 2). However, there is an apparent gap between the results of numerous tools and the formal requirements to deem a risk sufficiently mitigated or controlled. This gap between tools and abstract trustworthiness requirements should be addressed by future research. While testing methods need to

be further developed, a promising approach for concretizing requirements is to define (analogously to the protection profiles of the Common Criteria [30]) adequate classes of AI use cases (e. g., regarding the application environment or sector) and specify criteria as well as evaluation standards in more detail for the respective classes.

Acknowledgment: The authors would like to thank the consortium ZERTIFIZIERTE KI for the successful cooperation.

Funding: The development of this publication was supported by the Ministry of Economic Affairs, Innovation, Digitalization and Energy of the State of North Rhine-Westphalia as part of the flagship project ZERTIFIZIERTE KI (grant no. 005-2011-0048).

References

1. Angwin, J., J. Larson, S. Mattu and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.
2. Bogen, M. and A. Rieke. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity and Bias. Technical Report. Upturn.
3. Bouti, A. and D. A. Kadi. 1994. A state-of-the-art review of FMEA/FMECA. *International Journal of reliability, quality and safety engineering*, 1(04): 515–543.
4. Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, *et al*. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
5. Brundage, M., S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, *et al*. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.
6. Buolamwini, J. and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.
7. Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 2053951715622512.
8. Çınar, Z. M., A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael and B. Safaei. 2020. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. *Sustainability*, 12(19): 8211.
9. Crawford, K., R. Dobbe, T. Dryer, G. Fried, B. Green, *et al.* 2019. AI Now 2019 Report. New York: AI Now Institute.
10. Cremers, A. B., A. Englander, M. Gabriel, D. Hecker, M. Mock, *et al.* 2019. Trustworthy use of AI. Priorities from a philosophical, ethical, legal and technological viewpoint as a basis for certification of Artificial Intelligence. Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS, Sankt Augustin.
11. DIN e. V. and DKE 2020. Deutsche Normungsroadmap Künstliche Intelligenz.
12. Druzhkov, P. N. and V. D. Kustikova. 2016. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, 26(1): 9–15.
13. Dvijotham, K., R. Stanforth, S. Gowal, T. A. Mann and P. Kohli. 2018. A Dual Approach to Scalable Verification of Deep Networks. In *UAI (Vol. 1, No. 2, p. 3)*.
14. European Commission. 2018. Communication from the Commission, Artificial Intelligence for Europe (COM/2018/237 final).
15. European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts.
16. Floridi, L. and M. Taddeo. 2016. What is data ethics? *Philosophical Transactions of the Royal Society:* A.3742016036020160360. http://doi.org/10.1098/rsta.2016.0360.
17. Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, *et al.* 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4): 689–707.
18. Fredrikson, M., S. Jha and T. Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333.
19. Gladisch, C., C. Heinzemann, M. Herrmann and M. Woehrle. 2020. Leveraging Combinatorial Testing for Safety-Critical Computer Vision Datasets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 324–325.
20. Guegan, D. and B. Hassani. 2018. Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science*, 4(3): 157–171.
21. Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1): 99–120.
22. Hendrycks, D., N. Carlini, J. Schulman, and J. Steinhardt. 2021. Unsolved problems in ml safety. arXiv preprint arXiv:2109.13916.
23. High-level Expert Group on AI. 2019. Ethics Guidelines on Trustworthy AI. European Commission.
24. Houben, S., S. Abrecht, M. Akila, A. Bär, Brockherde, F., *et al.* 2021. Inspect, understand, overcome: A survey of practical methods for ai safety. arXiv preprint arXiv:2104.14235.
25. International Organization for Standardization. Standard ISO/IEC CD 23894. Information Technology – Artificial intelligence – Risk Management, under development.
26. International Organization for Standardization. Standard ISO/IEC CD 42001. Information Technology – Artificial intelligence – Management system, under development.
27. International Organization for Standardization. 2013. Standard ISO/IEC 27001:2013. Information technology – Security techniques – Information security management systems — Requirements.
28. International Organization for Standardization. 2015. Standard ISO 9001:2015. Quality management systems – Requirements.
29. International Organization for Standardization. 2015. Standard ISO 14001:2015. Environmental management systems — Requirements with guidance for use.

30. International Organization for Standardization. 2015. Standard ISO/IEC 15408-1:2009. Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model.
31. International Organization for Standardization. 2018. Standard ISO 31000:2018. Risk management – Guidelines.
32. ISO. 2021. The ISO Survey of Management System Standard Certifications – 2020 – Explanatory Note.
33. ISO. Management System Standards, n. d. URL: https://www.iso.org/management-system-standards.html (Accessed on 11.01.2022).
34. Jobin, A., M. Iencaand E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1: 389–399.
35. Katz, G., C. Barrett, D. L. Dill, K. Julian and M. J. Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, Cham.
36. Kersten, H., G. Klett, J. Reuter and K.-W. Schröder. 2020. *IT-Sicherheitsmanagement nach der neuen ISO 27001*. Springer Vieweg. ISBN 978-3-658-27691-1.
37. Killourhy, K. S. and R. A. Maxion. 2009. Comparing anomaly-detection algorithms for keystroke dynamics. *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pp. 125–134.
38. Lambert, G. 2017. A stroll down Quality Street. *ISOfocus 123 July-August 2017*. pp. 37–40.
39. Li, Z., X. Ma, C. Xu, and C. Cao. 2019. Structural coverage criteria for neural networks could be misleading. *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results*, pp. 89–92.
40. Liu, C., T. Arnon, C. Lazarus, C. Strong, C. Barrett and M. J. Kochenderfer. 2021. Algorithms for verifying deep neural networks. *Foundations and Trends® in Optimization*, 4(3–4): 244–404.
41. Mata, F. J., W. L. Fuerst and J. B. Barney. 1995. Information technology and sustained competitive advantage: A resource-based analysis. *MIS quarterly*, 487–505.
42. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.
43. Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.
44. Mock, M., A. Schmitz, L. Adilova, D. Becker, A. B. Cremers and M. Poretschkin. 2021. *Management System Support for Trustworthy Artificial Intelligence. Fraunhofer-Institut für Intelligente Analyse und Informationssysteme*. IAIS, Sankt Augustin.
45. Morley, J., L. Floridi, L. Kinsey and A. Elhalal. 2021. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Ethics, Governance, and Policies in Artificial Intelligence*. pp. 153–183. Springer, Cham.
46. OECD Council. 2019. Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).
47. OECD. AI, powered by EC OECD. 2021. Database of national AI policies. https://oecd.ai (Accessed on 29.03.2022).
48. Papernot, N., P. McDaniel, A. Sinha and M. Wellman. 2016. Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814.
49. Poretschkin, M., A. Schmitz, L. Adilova, M. Akila, D. Becker, *et al.* 2021. KI-Prüfkatalog: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS, Sankt Augustin.
50. Poretschkin, M., M. Mock and S. Wrobel. 2021. Zur Systematischen Bewertung der Vertrauenswürdigkeit von KI-Systemen. D. Zimmer (Hrsg.), *Regulierung für Algorithmen und Künstliche Intelligenz*.
51. Rostalski, F., and E. Weiss. 2021. Der KI-Verordnungsentwurf der Europäischen Kommission – Eine erste Analyse unter besonderer Berücksichtigung der Rolle von Zertifizierung. *Zeitschrift für Digitalisierung und Recht*, 4/2021.
52. Department of International Cooperation Ministry of Science and Technology (MOST), P.R.China. 2017. Next Generation Artificial Intelligence Development Plan. *China Science & Technology Newsletter No. 17*, Issued by State Council.
53. Sharma, R. and P. Yetton. 2003. The contingent effects of management support and task interdependence on successful information systems implementation. *MIS quarterly*, 533–556.
54. Toreini, E., M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya and A. Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 272–283.
55. Select Committee on Artificial Intelligence of the National Science & Technology Council. 2019. The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update.
56. Verma, S. and J. Rubin. 2018. Fairness definitions explained. *2018 IEEE/ACM international workshop on software fairness* (*fairware*), pp. 1–7.
57. Wachter, S., B. Mittelstadt and C. Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review*, 123: 735.
58. Whitman, M. E. and H. J. Mattord. 2021. *Principles of information security*. Cengage learning.
59. Yu, K. H., A. L. Beam and I. S. Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10): 719–731.
60. Yurtsever, E., J. Lambert, A. Carballo and K. Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8: 58443–58469.
61. Zhang, H., T. W. Weng, P. Y. Chen, C. J. Hsieh and L. Daniel. 2018. Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems* 31.
62. Zhang, J. M., M. Harman, L. Ma and Y. Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*.
63. Zhang, Y., P. Tiňo, A. Leonardis and K. Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*.
64. Zhou, J., F. Chen, A. Berry, M. Reed, S. Zhang and S. Savage. 2020. A Survey on Ethical Principles of AI and Implementations. 2020 *IEEE Symposium Series on Computational Intelligence (SSCI)*, 3010–3017.

Bionotes

Anna Schmitz
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany
anna.schmitz@iais.fraunhofer.de

Anna Schmitz is a scientific employee at the Fraunhofer Institute for Intelligent Analysis and Information Systems. Her research focuses on the topics of trustworthy AI and certification of AI applications and is primarily conducted within the interdisciplinary project ZERTIFIZIERTE KI. Currently, she is developing systematic approaches for risk analysis and evaluation of AI systems, which have contributed to the 'IAIS KI-Prüfkatalog', for example, and which are being applied in pilot AI assessments with industry clients. Anna Schmitz studied mathematics at the Universities of Cologne and Cambridge.

Maram Akila
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany
ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany
maram.akila@iais.fraunhofer.de

Dr. Maram Akila conducts research at the Fraunhofer Institute for Intelligent Analysis and Information Systems on safeguarding and certification of complex ML systems. One focus within the project 'KI-Absicherung' is on methods in the field of autonomous driving, especially object detection. Since the beginning of 2022, he has also been supporting the ML2R competence center as coordinator for "Trustworthy ML". Dr. Akila studied physics at University Duisburg Essen and obtained his doctorate in theoretical quantum physics on the dynamics of chaotic many-body systems.

Dirk Hecker
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany
ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany
dirk.hecker@iais.fraunhofer.de

Dr. Dirk Hecker is Deputy Director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and Managing Director of the »Fraunhofer Big Data and Artificial Intelligence Alliance« which is made up of more than 30 Fraunhofer institutes working in cross-sector Big Data research and technology development. He has many years experience in managing Data Mining and Machine Learning research and industry projects. His current focus of activity is in Big Data Analytics, Predictive Analytics, Deep Learning and Mobility Mining. Dr. Hecker studied geoinformatics at the universities of Cologne and Bonn; he obtained his doctorate at the University of Cologne.

Maximilian Poretschkin
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany
ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany
Institut für Informatik III, Rheinische-Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
maximilian.poretschkin@iais.fraunhofer.de

Dr. Maximilian Poretschkin is Senior Data Scientist and Team Lead at the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and is responsible for activities in the area of validation, testing and standardization of artificial intelligence. His research focuses on the development of testing principles and tools that can be used to assess whether AI systems meet recognized quality standards. In addition, he is co-author of one of the first testing catalogs for AI systems. Currently, Dr. Poretschkin leads the consortium ZERTIFIZIERTE KI and the working group "Testing and Certification" of the German Standardization Roadmap. Dr. Poretschkin studied Physics in Bonn and Amsterdam.

Stefan Wrobel
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany
ML2R – Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr, Dortmund, Germany
Institut für Informatik III, Rheinische-Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
stefan.wrobel@iais.fraunhofer.de

Prof. Dr. Stefan Wrobel is Professor of Computer Science at University of Bonn and Director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS. He studied computer science and artificial intelligence in Bonn and Atlanta, Georgia/USA (M.S., Georgia Institute of Technology) and obtained his PhD at the University of Dortmund. After holding positions in Berlin and Sankt Augustin he was appointed Professor of Practical Computer Science at Magdeburg University, before taking up his current position in 2002. Professor Wrobel's work is focused on questions of the digital revolution, in particular intelligent algorithms and systems for the large-scale analysis of data and the influence of Big Data/Smart Data on the use of information in companies and society.