electronics

MDPI

Article

Trustworthiness Optimisation Process: A Methodology for Assessing and Enhancing Trust in AI Systems

Mattheos Fikardos [1,*], Katerina Lepenioti [1], Dimitris Apostolou [2] and Gregoris Mentzas [1]

[1] Information Management Unit, Institute of Communication and Computer Systems (ICCS), School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), 10682 Athens, Greece; klepenioti@mail.ntua.gr (K.L.); gmentzas@mail.ntua.gr (G.M.)
[2] Department of Informatics, University of Piraeus, 18534 Pireas, Greece; dapost@unipi.gr
* Correspondence: mfikardos@mail.ntua.gr

Abstract: The emerging capabilities of artificial intelligence (AI) and the systems that employ them have reached a point where they are integrated into critical decision-making processes, making it paramount to change and adjust how they are evaluated, monitored, and governed. For this reason, trustworthy AI (TAI) has received increased attention lately, primarily aiming to build trust between humans and AI. Due to the far-reaching socio-technical consequences of AI, organisations and government bodies have already started implementing frameworks and legislation for enforcing TAI, such as the European Union's AI Act. Multiple approaches have evolved around TAI, covering different aspects of trustworthiness that include fairness, bias, explainability, robustness, accuracy, and more. Moreover, depending on the AI models and the stage of the AI system lifecycle, several methods and techniques can be used for each trustworthiness characteristic to assess potential risks and mitigate them. Deriving from all the above is the need for comprehensive tools and solutions that can help AI stakeholders follow TAI guidelines and adopt methods that practically increase trustworthiness. In this paper, we formulate and propose the Trustworthiness Optimisation Process (TOP), which operationalises TAI and brings together its procedural and technical approaches throughout the AI system lifecycle. It incorporates state-of-the-art enablers of trustworthiness such as documentation cards, risk management, and toolkits to find trustworthiness methods that increase the trustworthiness of a given AI system. To showcase the application of the proposed methodology, a case study is conducted, demonstrating how the fairness of an AI system can be increased.

Keywords: trustworthy AI; artificial intelligence; fairness; trustworthiness; AI lifecycle; machine learning; trust in AI

Academic Editor: Domenico Rosaci

Received: 26 February 2025
Revised: 28 March 2025
Accepted: 31 March 2025
Published: 3 April 2025

Citation: Fikardos, M.; Lepenioti, K.; Apostolou, D.; Mentzas, G. Trustworthiness Optimisation Process: A Methodology for Assessing and Enhancing Trust in AI Systems. *Electronics* **2025**, *14*, 1454. https://doi.org/10.3390/electronics14071454

1. Introduction

The need for trustworthy artificial intelligence (TAI) has become increasingly critical, as AI systems are being integrated into decision-making processes across domains such as health care, finance, and law enforcement [1]. Trustworthy AI is essential for ensuring ethical and unbiased outcomes, as unregulated and poorly designed systems can perpetuate social inequalities and harm individuals and communities. As Baeza-Yates and Fayyad [2] argue, the rapid adoption of AI without adequate safeguards can exacerbate existing societal inequities, perpetuate bias, and erode public trust. Furthermore, fostering trustworthiness is not merely a technical challenge, but also a socio-political imperative, necessitating the integration of ethical guidelines and regulatory frameworks to guide AI development responsibly.

Trustworthiness is a multi-dimensional construct that integrates ethical principles, technical solutions, and user-centric designs. By embedding these elements into AI systems, society can leverage AI's transformative potential while minimising risks such as misinformation and algorithmic discrimination [3]. Creating trustworthy AI is essential for regulatory compliance and long-term societal benefit. Díaz-Rodríguez et al. [4] highlight the importance of connecting ethical principles to actionable requirements, such as accountability measures and technical standards. These connections ensure that AI systems are not only theoretically aligned with ethical values, but also practically implemented to address real-world challenges.

There are already many methods, algorithms, and techniques designed to address various dimensions of trustworthiness. The literature reflects a proliferation of these tools, each aimed at specific facets such as fairness, explainability, and robustness. For example, Prem [5] emphasised the persistent gap between high-level ethical principles and their practical implementation and reviewed frameworks that provide a structured approach to evaluating and enhancing AI systems. Narayanan and Schoeberl [6] introduced a selection matrix to help AI practitioners identify approaches most suitable for their specific needs, showcasing the diverse range of approaches available. For instance, in the case of fairness, algorithms and techniques to mitigate bias in AI systems include pre-processing techniques such as re-weighting and re-sampling aim to balance datasets [7], in-processing approaches like adversarial debiasing [8], and post-processing methods such as calibration and equalised odds [9].

Nevertheless, there is still a need for a concrete, procedural method that will help AI decision-makers, scientists, and practitioners select and apply the alternative methods in a consistent manner across the AI lifecycle. The present paper aims to fill this gap by proposing a method that facilitates the selection of improving AI trustworthiness while meeting ethical and societal expectations. In this paper, we introduce a structured, risk-based methodology for assessing the trustworthiness characteristics of an AI system that makes explicit reference to various algorithms and techniques for improving TAI. Our method makes use of the recent documentation trend in AI system development that uses tools such as data cards and model cards to enhance transparency and accountability. Such documentation tools provide standardised formats for describing critical details about datasets and models, including their design, intended use, and potential limitations. We extend the concept of these cards to introduce the method cards, which take into account the alternative algorithms for TAI improvement across the AI lifecycle.

The structure of the paper is as follows. Section 2 provides related work in the field of TAI, highlighting related concepts and challenges. Section 3 presents the proposed methodology, detailing its stages and their functionality. In Section 4, a case study is conducted to showcase the usage of the proposed methodology. Finally, Section 5 concludes the article with a short discussion of the results and future directions.

2. Related Work

2.1. Procedural Methodologies for TAI

As artificial intelligence (AI) increasingly influences diverse domains, ensuring the trustworthiness of AI systems has become paramount. Trustworthy AI encompasses ethical principles, such as fairness, transparency, and accountability, operationalised into actionable processes. While numerous technical algorithms address these principles, procedural methodologies provide the critical frameworks and processes needed to systematically evaluate AI systems for compliance with ethical and societal standards. The literature reflects a proliferation of such methods, each tailored to address specific aspects of trustworthiness. For instance, Prem [5] reviewed over 100 AI frameworks, emphasising the gap between

theoretical principles and practical tools. Similarly, Boza and Evgeniou [10] highlighted the need for structured processes to translate high-level guidelines into implementation practices, while Narayanan and Schoeberl [6] introduced a matrix for selecting appropriate frameworks, underscoring the diversity in approaches. Below, we explore certain key procedural methodologies, outlining their contributions and applications.

The CapAI methodology, developed by Floridi et al. [11], is aligned with the European Union's Artificial Intelligence Act. It offers a conformity assessment procedure designed to ensure that AI systems meet ethical and legal standards. CapAI views AI systems across the entire AI lifecycle from design to retirement. It defines and reviews current practices and enables technology providers and users to develop ethical assessments at each stage of the AI lifecycle. The procedure consists of an internal review protocol, an external scorecard, and a summary data sheet. The process involves auditing systems across three layers: design, implementation, and governance [12]. Ethics-based auditing is integral to this method, providing a structured evaluation of AI systems in practice [13]. Its modular approach makes it adaptable across various sectors.

Another method with many applications is Z-Inspection®, developed by Zicari et al. [14]. Z-Inspection® employs a multi-stage evaluation process to assess the trustworthiness of AI systems. Originating in health care contexts, it integrates stakeholder involvement, contextual analysis, and ethical review. The method has been applied to systems like COVID-19 diagnostic tools and skin lesion classifiers, demonstrating its practical relevance [15]. By incorporating stakeholder feedback and real-world testing, Z-Inspection ensures the alignment of AI systems with ethical norms and societal expectations.

Poretschkin et al. [16] introduced the AI Assessment Catalog, a guideline for evaluating trustworthy AI. This methodology provides a comprehensive checklist covering aspects like privacy, robustness, and ethical alignment. It emphasises scalability and flexibility, allowing organisations to adapt the assessment process to specific use cases. The catalog bridges the gap between theoretical principles and operational practices.

Incorporating ethics in the AI development lifecycle is another approach. For instance, ECCOLA, proposed by Vakkuri et al. [17], emphasises embedding ethical considerations and provides a structured toolkit to ensure ethical alignment throughout design, implementation, and deployment. By integrating ethics into every phase of development, ECCOLA ensures that trustworthiness is not an afterthought, but a foundational attribute of AI systems. Another approach is the Set–Formalize–Measure–Act (SFMA) methodology, a pragmatic framework for operationalising trustworthiness, proposed by Nasr-Azadani and Chatelain [18]. SFMA involves setting ethical goals, formalising them into measurable metrics, evaluating system performance, and taking corrective action. This iterative approach ensures continuous improvement and alignment with evolving ethical standards.

Brunner et al. [19] proposed a comprehensive framework for ensuring AI trustworthiness through based assessments. It emphasises the identification and mitigation of risks throughout the AI lifecycle. By integrating technical and procedural safeguards, this framework ensures that systems are both reliable and aligned with societal values. Baker-Brunnbauer [20] proposed the TAII framework, which offers a structured approach for implementing trustworthy AI systems. TAII focuses on transparency, accountability, and interpretability, providing a roadmap for embedding these principles into the AI development process. Applications, such as its influence on robotics design, demonstrate TAII's utility across domains [21].

Baldassarre et al. [22] introduced the POLARIS framework to guide organisations in developing trustworthy AI systems. It integrates ethical considerations, regulatory compliance, and stakeholder engagement into a cohesive assessment process. Iterative evaluations ensure adaptability to changing contexts and requirements. The OOD-BC methodology by

Stettinger et al. [23] focuses on assessing trustworthiness in high-risk applications, such as autonomous systems. By emphasising the fact that the Operational Design Domain (ODD) concept and its associated behavioural competencies play a pivotal role, this approach ensures that AI systems maintain reliability under diverse and unpredictable conditions.

In addition to research in assessing all trustworthiness dimensions, some methodologies focus on specific dimensions. For example, Confalonieri and Alonso-Moral [24] introduce an operational framework for guiding human evaluation in explainable and trustworthy AI. The framework focuses on the human-centric assessment of AI explanations, emphasising transparency, comprehensibility, and actionable insights, and provides structured guidelines to evaluate the usability and cognitive impact of AI-generated explanations across diverse contexts. Their methodology integrates qualitative and quantitative evaluation metrics, addressing both subjective and objective criteria in order to ensure that AI explanations align with user needs and ethical standards.

There have also been some methods that aim to incorporate trustworthiness aspects during the AI system development process. For instance, Hohma and Lütge [25] explore the translation of abstract principles of AI trustworthiness into a concrete development process for AI systems. They propose a framework that combines normative principles, such as fairness and transparency, with practical methodologies for ensuring their application, and highlight the significance of stakeholder involvement, iterative testing, and accountability mechanisms as integral components. In a similar fashion, Ronanki et al. [26] focus on requirements engineering as a central pillar for developing trustworthy autonomous systems and present a set of recommendations to guide the development process, emphasising systematic identification and prioritisation of trustworthiness attributes, such as safety, reliability, and ethical compliance. Their approach integrates stakeholder perspectives, iterative refinement, and traceability to ensure alignment between design objectives and societal expectations.

Finally, many industry leaders like Microsoft and IBM, consultancies like Accenture and PwC, and specialised companies like Digital Catapult have developed methodologies to operationalise AI ethics. For instance, Microsoft [27] bases its approach on the framework of the National Institute of Standards and Technology, while PwC's Responsible AI framework focuses on governance and compliance, offering tools to manage risks throughout the AI lifecycle; see [28]. Digital Catapult [29] emphasises embedding ethics in the design phase, while IBM [30] outlines the use of WatsonX Governance to provide automated tools for monitoring compliance with trustworthiness standards. In addition to its own method outlined in Accenture [31], Accenture has partnered with Amazon Web Services to provide an integrated toolkit for responsible AI [32].

Procedural methodologies are indispensable for assessing the trustworthiness of AI systems. By combining AI trustworthiness principles with actionable processes, these methods enable organisations to navigate the complex challenges of AI governance. From conformity assessments like CapAI to industry-specific approaches like OOD-BC and ECCOLA, each methodology contributes to a robust ecosystem for building trustworthy AI.

However, despite providing valuable approaches, they are not coupled to the multitude of existing algorithms and toolkits that are available for the assessment of the various dimensions of trustworthiness like fairness, privacy, explainability, etc. In this paper, we aim to develop such links and integrate diverse approaches to support trustworthiness assessment and improvement in a comprehensive manner.

2.2. TAI Documentation

The inherent complexity and opacity of AI systems often hinder stakeholders, such as developers, users, and policymakers, from fully understanding their functionality and

implications. This lack of transparency exacerbates challenges related to trust, fairness, accountability, and compliance with ethical and legal standards [33]. Documentation has emerged as a key solution to address these concerns, with frameworks tailored to providing structured, accessible, and comprehensive descriptions of AI systems and their components. Various approaches have been proposed in the literature, each with unique characteristics and purposes, reflecting the evolving landscape of AI governance [34].

As Oreamuno et al. [35] noted, third-party datasets and models often lack sufficient documentation, creating barriers to reproducibility, risk assessment, and informed decision-making. The increasing regulation of AI systems, such as the European AI Act, further necessitates documentation that bridges technical detail and policy requirements. In the following, we review prominent approaches to AI documentation, focusing on the various specific types of "cards" as practical tools for transparency: data cards, model cards, system cards, use case cards, and integrated cards.

Data cards offer structured documentation of datasets, focusing on their provenance, composition, and intended use. Pushkarna, Zaldivar, and Kjartansson [36] introduced data cards to promote responsible AI development by highlighting potential biases, ethical concerns, and technical limitations inherent in datasets. Data cards typically include details on collection methods, pre-processing, licensing, and known risks. By fostering informed decision-making, these cards aim to mitigate issues stemming from poorly understood and opaque datasets.

Model cards, pioneered by Mitchell et al. [37], provide a standardised format for documenting machine learning models. They outline key attributes such as intended use cases, performance metrics, limitations, and ethical considerations. Building on this foundation, Crisan et al. [38] proposed interactive model cards, enhancing user engagement through dynamic visualisations and detailed metadata. Recent large-scale studies, such as Liang et al. [39], highlight the proliferation of model cards across thousands of AI systems, showcasing their utility while identifying gaps, such as inconsistencies in reporting depth.

System cards, developed by Meta, extend the documentation paradigm to encompass entire AI systems rather than individual components. As described by Alsallakh et al. [40], system cards provide an overarching view of how AI systems function, interact, and impact users. Examples include Meta's cards for Facebook and Instagram, which explain algorithms governing content recommendations [41]. However, these efforts have faced criticism, with the Mozilla Foundation arguing that Meta's approach often lacks substantive detail and fails to address transparency comprehensively [42].

Documenting the data and the models used in an AI system misses the intricate details and the context of the specific use case, which may be quite relevant for assessing the trustworthiness of the AI system (e.g., the requirements of a human-resource management system are quite distinct from those of a medical system). The need to address the specificities of the system's usage is covered by use cards, introduced by Hupont et al. [43], who proposed use case cards to document AI applications tailored to specific scenarios. These cards include contextual information, risks, and ethical considerations relevant to particular applications.

Finally, integrated cards aim to consolidate all the above diverse documentation needs into a unified framework. Gursoy and Kakadiaris [44] and Golpayegani et al. [45] advocate for machine-readable formats that harmonise data, model, and system-level documentation, facilitating compliance with regulatory frameworks like the EU AI Act. These cards enable stakeholders to trace decision-making process.

Recent research trends emphasise leveraging large language models (LLMs) to enrich documentation practices. For instance, Giner-Miguelez et al. [46] explore using LLMs to generate contextual metadata for datasets, enhancing their interpretability. However,

studies such as Yang et al. [47] reveal persistent gaps in dataset and model documentation, including incomplete and outdated information. Mehta et al. [48] highlight the need for dynamic documentation approaches that adapt to the evolving nature of AI systems.

This multitude of alternative approaches to documentation has led to a recent effort to identify best practices and provide concrete guidelines for effective documentation. The CLeAR framework, proposed by Chmielinski et al. [49], outlines best practices for AI documentation, emphasising clarity, legibility, and relevance. This framework serves as a roadmap for developing documentation that aligns with ethical principles and regulatory requirements.

The proliferation of documentation approaches reflects the critical role of transparency in AI development. While tools like data cards, model cards, and system cards cater to specific aspects of AI systems, integrated cards and frameworks like CLeAR offer holistic solutions.

Challenges remain, however, particularly in incorporating information about the available algorithms and techniques for assessing and improving trustworthiness, ensuring comprehensiveness, consistency, and adaptability.

2.3. Trustworthy AI Pillars

Achieving AI trustworthiness is a highly complex task due to the broad variability in AI systems and social-technical environments in which they can be deployed [50], the multitude of relevant trustworthiness characteristics (TC) [51–53], and the range of disciplines in which trustworthy AI has been addressed [54]. As noted by Li et al. [55], AI trustworthiness is the result of the combined effect of relevant trustworthiness characteristics, which cannot be assessed and tackled by considering single AI lifecycle stages in isolation [53]. Before presenting the proposed methodology, we outline the key pillars of trustworthy AI, as depicted in Figure 1, and explain why these are important foundations for the methodology. The pillars include the lifecycle dimension, the assets, the actors, risk management, as well as the toolkits and algorithms.

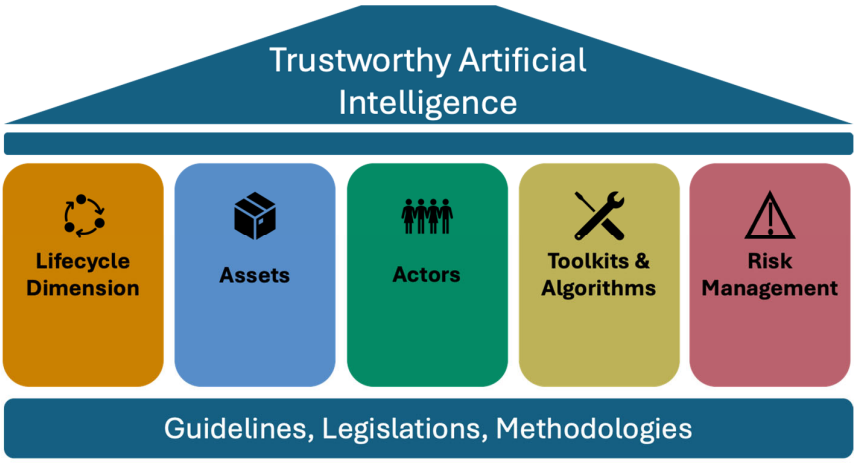

Figure 1. Trustworthy artificial intelligence pillars.

Lifecycle Dimension. The AI lifecycle manages the complicated process of moving from a business problem to an AI solution that solves that problem, ensuring the expected business results are attained. The process comprises various tasks and decisions that drive the development and deployment of the AI solution and is often iterative, requiring steps in the lifecycle to be revisited throughout the design, development, and deployment stages [56]. The high-level stages of the lifecycle are considered to be (i) planning and requirements, (ii) designing, (iii) development, (iii) deployment, and (iv) operation and

monitoring. Deviations from these phases can be found in the literature. OECD [57] and NIST [53] differentiate the development phase into the stages of (i) building and using the models and (ii) verifying and validating them. ENISA [52] uses a more detailed view of the lifecycle with 11 distinct stages that fall into the high-level phases. Kaur et al. [54] and Calegari et al. [58] follow a similar approach, with differences in the naming of some stages, such as from Modeling to Development and from Oversight to Planning. Some approaches focus more on the ML lifecycle [59–61] or AI product lifecycle [55] and follow the same phases as the system lifecycle.

The initial and final stages of the lifecycle are typically business-centric; the requirements and objectives of the AI system are defined based on business needs, and the business impact is monitored. Data play a pivotal role throughout the AI system lifecycle, serving as the foundation of the system that influences its feasibility and accuracy. For this reason, the design and development stages are characterised as data-intensive and data-centric, which occur early during the lifecycle. Following the data-related activities and processes, the lifecycle progresses to the model-centric stages, which mainly relate to the development and deployment of the AI system and, specifically, its AI models.

Derived from the AI model-related stages, another set of characteristics that focuses on the processing of the data can be used. Based on the status of data processing, the lifecycle stages can be categorised into three groups: (i) pre-processing (before the data are used as input to the AI models); (ii) in-processing (during data processing by AI models), and (iii) post-processing (after AI model deployment) [62]. Another important characteristic vital to the perspective of trustworthiness is human involvement. This spans from the planning and creation of system expectations and requirements (human-before-the-loop), through the development and deployment stages of the lifecycle (human-in-the-loop), to the operation and monitoring of the system for oversight (human-over-the-loop) [54].

Assets. The assets [52] and dimensions [53] describe the aspects or entities of focus during one or multiple stages of the lifecycle. For example, the assets could be the application domain, data, AI models, artifacts, and processes that constitute the AI system and directly affect and shape its trustworthiness. On the other hand, from the perspective of AI trustworthiness, these assets should also reflect the trustworthiness characteristic of interest, such as robustness and fairness, while being linked to every stage of the lifecycle where applicable.

Actors. Several stakeholders, i.e., AI actors, are involved in the AI lifecycle. Depending on the current stage of the AI system, different domain experts, policymakers, users, and stakeholders may contribute or be responsible. Depending on the application domain, the number and variety of assets and AI actors can influence the lifecycle of the AI system. A comprehensive analysis of the different AI actors' categories and their respective tasks can be found in Appendix A of the NIST [53] AI Risk Management Framework (AI RMF), where several actors are identified in various tasks, such as the AI design, development, deployment, operation, monitoring, testing, evaluation, verification, and validation.

Toolkits and Algorithms. Important enablers of AI trustworthiness are the practical toolkits, methods, and algorithms that can be employed to monitor and mitigate trustworthiness issues. Mentzas et al. [63] surveyed the prominent toolkits, both commercial and academic. They outlined the underlying methods and algorithms that measure the performance of the AI system with respect to specific trustworthiness characteristics, as well as those that make changes to the AI system to mitigate identified trustworthiness shortcomings. In the rest of the paper, we refer to the former as metrics and to the latter as trustworthiness methods. For example, algorithmic implementations of fairness criteria such as equalised odds and equal opportunity [9] can be used to provide measurements regarding predictive discrimination between sensitive variables. On the other hand, methods

such as adversarial training [64] can be employed to enhance the training of models and mitigate limitations regarding their robustness.

Risk Management. Organisations such as ENISA, NIST, and OECD developed frameworks and reports for practising trustworthy and responsible AI development, where risk management principles are incorporated. The AI RMF from NIST [53] aims to promote the trustworthy development and use of AI systems and assist organisations in managing the associated risks. The framework is built on top of four core functions (map, measure, manage, and govern) and considers the environment, AI actors, and TC of the AI system. ENISA's framework for AI cybersecurity practices [52] provides guidelines for stakeholders to secure AI systems throughout their lifecycle with practices such as risk assessments and management. In addition, the OECD issued a report for governing and managing the risk of AI from an accountability perspective [65]. The report advocated embedding risk management processes throughout the AI system lifecycle to enable organisations to manage AI risk with a four-step process (define, assess, treat, and govern). These frameworks are aligned with the AI Act [66] where AI systems have to be categorised into four levels of risk. Based on the above, the AI system's limitations in terms of trustworthiness can be framed as risks that need to be dealt with, where a risk represents the probability (likelihood) and consequence magnitude (impact) of a positive or negative event that can occur [53]. Moreover, the associated risk can be linked to the compromise of specific TCs (e.g., fairness, transparency, robustness) and be assessed with specific processes and measures [65].

2.4. Challenges

Despite the significant progress in the fields of TAI with emerging guidelines, procedural methodologies, regulatory frameworks, and technical approaches for assessing and mitigating trustworthiness issues of AI and the variety of problems these innovations address, there are still shortcomings and challenges that need to be addressed. In this section, we cite some of those challenges that are related to our work.

Lifecycle Perspective. Research developments in TAI are ahead of their practical utilisation, thus creating a gap [54]. This is also related to limitations for evaluating trustworthiness [55] and is further exaggerated by the current state of the AI lifecycle. Our analysis of the literature reveals that the trustworthiness perspective is often addressed through the AI lifecycle, but not in a systematic and rigorous manner, and without being considered an independent procedure. As a consequence, the trustworthiness of the AI system and underlying AI model cannot be performed in a single step; it needs to be developed and maintained from a lifecycle perspective. We argue that an independent, yet strongly related lifecycle for the trustworthiness perspective should be considered. Both perspectives of the lifecycle should cooperate synchronised and in harmony, with each one addressing its domain, but also supplementing the other's limitations and weaknesses. The trustworthiness lifecycle should be flexible and expandable to incorporate the aforementioned dimensions and characteristics while also being able to be adapted to the various phases of the AI lifecycle.

Trustworthiness Characteristics and Conflicts. Conflicts between the various dimensions of trustworthiness (e.g., accuracy vs. fairness) have been identified as a challenge in recent surveys [55,62], and their exact relationships are yet to be fully investigated. Satisfying trustworthiness requirements may entail opposing requirements, and the mitigation of one characteristic may have adverse effects on others [53,67]. This has created the need to comprehensively consider their contradicting effects when enforcing trustworthiness. Despite the absence of a comprehensive methodology that addresses conflicts and reconciliation between all the TCs, researchers experimented with multi-criteria approaches to assess and evaluate trustworthiness criteria to enhance decision-making procedures [68–70].

Human Oversight and Collaboration. Another major challenge is the relationship between humans and the trustworthiness procedures. From the AI system analysis, it is clear that AI actors from multiple disciplines must contribute to the trustworthiness process. Further, the AI actors should be fully utilised and included throughout the lifecycle, giving the human a protagonist role, not only in all the stages of AI system development, but also in all the stages of trustworthiness assessment and enhancement. This indicates that humans are educated and informed about TAI and that their expertise and preferences are considered. They actively participate in and supervise the related procedures that operationalise or even automate, to some extent, the trustworthiness of the AI systems.

3. Research Method

The proposed Trustworthiness Optimisation Process (TOP) aims to equip practitioners and AI actors with a methodology that operationalises the enforcement of trustworthiness. To comprehensively design the proposed approach, we followed the four-step research method depicted in Figure 2.

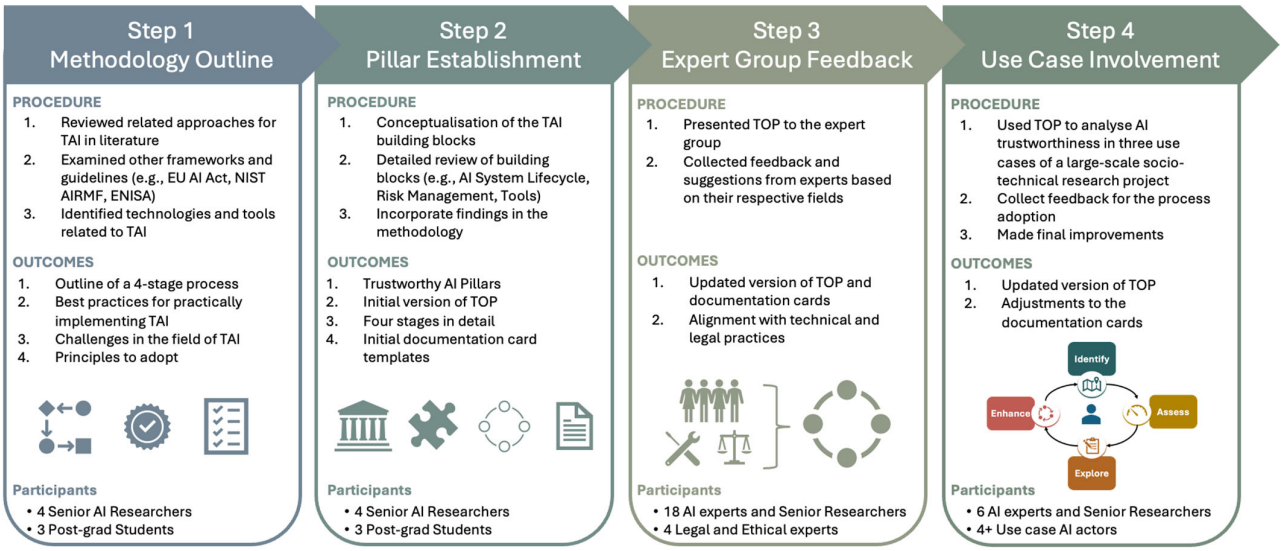

Figure 2. Design procedure of TOP.

In the first step, we reviewed the field of TAI and specifically investigated how the trustworthiness of an AI system can be optimized. This led us to survey the state-of-the-art (SOTA) approaches, frameworks, and guidelines in the field, as well as related technologies and toolkits. Drawing inspiration from our findings, we outlined a process consisting of four necessary stages to effectively assess and enhance the trustworthiness of an AI system while incorporating the best implementing practices. The review analysis indicated some shortcomings of current approaches, which have been grouped into some key challenges (Section 2.4). To address these challenges, the proposed process was designed based on the following five desired properties.

- **Vertical compatibility** with the AI lifecycle, enabling application across all stages;
- **Extensibility,** facilitating continuous enrichment of the available trustworthiness methods pool;
- **Conflict consideration** to tackle friction and negative implications between the TC;
- **Human-in-the-centre** approach, placing humans as the focal point and including them through the process;
- **Multidisciplinary engagement,** where multiple stakeholders participate and provide inputs where needed.

Next, we investigated the underlying building blocks that are necessary for the trustworthiness optimisation. Each building block was analysed further and incorporated into the four-stage process. The identified building blocks yielded the trustworthy AI pillars (Section 2.3) and resulted in the initial version of TOP with detailed stages and templates for the documentation cards.

We then elicited feedback from experts in the field. For this reason, TOP was presented to a group of multidisciplinary experts. A total of 22 experts were consulted, including individuals with backgrounds in AI research, technical and algorithmic implementation, risk management, business decision-making, law, and ethics. Taking into account the feedback and suggestions from the experts, we updated TOP and aligned it with technical and legal practices. For instance, the feedback emphasised the importance of TCs, particularly from the perspective of risk management, as well as their connection with trustworthiness methods that are split between assessment and enhancement methods.

Finally, we used TOP as the foundation for analysing the trustworthiness in three real-life use cases of a large-scale socio-technical research project. The involved use cases—comprising a maritime port, a media institution, and a medical hospital—were represented by four key representatives who served as AI actors for their respective AI systems. TOP was used to design the conceptual architecture and the user journeys of the research project. From the use cases, we were able interact with a diverse set of AI systems and requirements. The maritime port used an XGBoost algorithm for time series prediction with tabular datasets (numeric and categorical) and was more interested in the TCs of fairness and explainability. The media institution incorporated large language models for fake news and disinformation perditions and was more interested in the TCs of fairness and accuracy. Lastly, the medical hospital used multiple AI models (including image classification) related to the patient treatment plan, where the TCs of fairness, robustness, and accuracy were important. These differences between the use cases contributed to designing a generic process that can be employed in varying settings. The collected feedback was used to make the final improvements, resulting in the final version of TOP and adjustments to the documentation cards.

4. Methodology

In this section, we present the final version of TOP, which augments the AI system lifecycle and guides the process of optimising its trustworthiness with the capability to be adapted to meet requirements and incorporate metrics, methods, and algorithms to effectively assess and mitigate risks.

As depicted in Figure 3a, the proposed methodology is comprised of four stages: Identify, Assess, Explore, and Enhance. Initially, in the Identify stage, information related to the societal and technical environment of the AI system is gathered, recorded in the form of documentation cards, and made accessible to the following stages. In the second stage, Assess, the quantitative and risk assessment of the system is executed by leveraging existing metrics and frameworks. In the third stage, Explore, a search for optimal solutions that mitigate trustworthiness risks is executed. These solutions are constructed based on the plethora of existing technical algorithms and toolkits that mitigate limitations related to trustworthiness. In the final stage, Enhance, a solution that meets the current requirements of the AI system is selected and implemented in the AI system through its lifecycle. At the same time, it is continually monitored to ensure the preservation of its performance.

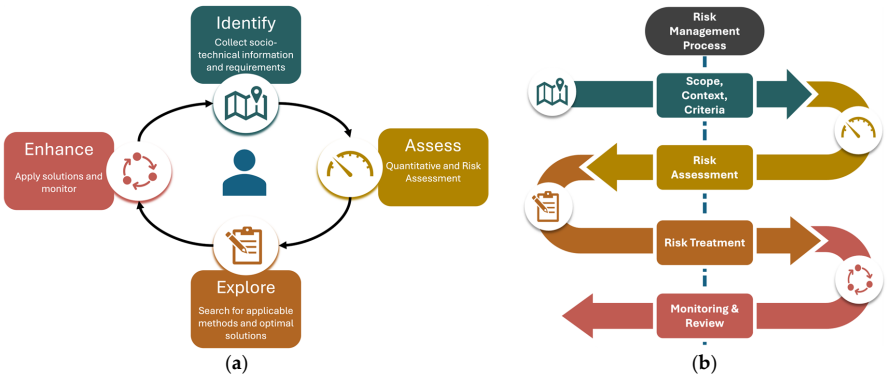

Figure 3. Trustworthiness Optimisation Process overview. (**a**) The four stages of TOP in sequence, highlighting their main functionality; (**b**) interplay between TOP and the risk management process.

Moreover, as depicted in Figure 3b, TOP was designed to follow the risk management process of ISO 31000 [71]. The process follows the steps of (i) Scope, Context, and Criteria, (ii) Risk Assessment, and (iii) Risk Treatment, while it is continually Communicated, Monitored, and Reported. These steps coincide with the stages of TOP: The context establishment occurs during the Identify stage, while the Risk Assessment and Treatment materialize in the Assess and Explore stages, respectively. The ultimate Monitoring and Reviewing take place at the Enhance stage, even though intermediary reports and monitoring can occur through the process of communicating important information. With the above design, TOP aims to be integratable with risk management frameworks in the field of TAI through the common process, where the principles and framework can be adjusted accordingly. In this setting, TOP also adheres to standards focused on AI management such as the ISO 4200 and ISO 23894 [72,73], as both follow the ISO 31000 guidelines [71].

The stages of TOP progress sequentially, with each depending on the output of the previous one and each having a specific goal and outcome that can be executed at any point of the AI system lifecycle based on the available information. The remainder of this section provides a formulation of the problem TOP aims to solve and a detailed description of its stage, as shown in Figure 4.

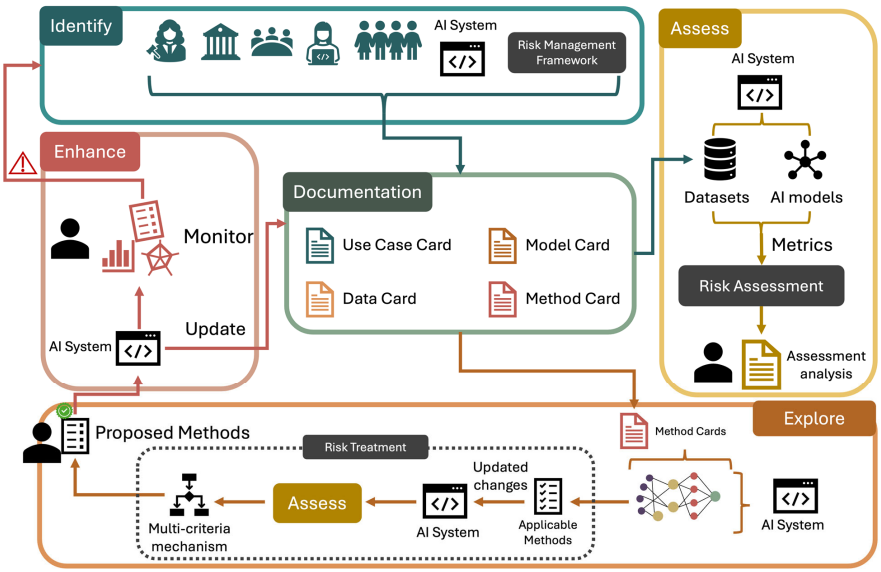

Figure 4. Stages of the Trustworthiness Optimisation Process.

4.1. Problem Definition

The problem the proposed process tries to address is the trustworthiness assessment of an AI system, followed by the identification and implementation of trustworthiness

enhancement methods that effectively mitigate its associated risks, while also integrating the aforementioned pillars. The formulation is defined as follows.

Let the AI system (AIS) be a set of stages, assets, actors, and risks:

$$AIS =< S_{i=1}^{i=n}, A_{j=1}^{j=m}, H_{k=1}^{k=l}, R_{p=1}^{p=q} > \tag{1}$$

where $i,\ j,\ k,\ p \in \mathbb{N}$, and $n,\ m,\ l,\ q$ are the number of stages (S), assets (A), AI actors (H), and risks (R) of the AI system, respectively. For instance, an AIS can be compromised by a set of A that includes the different datasets and AI models of the system. The different H are the domain experts, data analysts, and engineers who work on the system and the set of associated R with regards to trustworthiness.

Define TCs as the set of f trustworthiness characteristics that need to be preserved:

$$TCs = \Big\{\ TC_1,\ TC_2,\ \ldots\ TC_f,\ \Big\},\ f \in \mathbb{N} \tag{2}$$

and the available assessment methods (i.e., metrics) am as:

$$am = \{am_1,\ \ldots,\ am_e\},\ e \in \mathbb{N}, \text{where } am_e\Big(TC_f,\ AIS\Big) = T \tag{3}$$

where T is the technical (or quantitative) assessment of the AIS. Thus, the trustworthiness assessment TA of the system can be calculated by incorporating T and a risk management framework RM with the system risks R^{AIS} to produce the overall assessment of the system as:

$$TA = RM\Big(R^{AIS},\ T\Big) \times T \tag{4}$$

Let the applicable enhancement methods (i.e., trustworthiness methods) em be:

$$em = \{em_1, \ldots, em_z\}, z \in \mathbb{N} \tag{5}$$

where $em_z\Big(TCs_f,\ AIS,\ TA\Big) =<\ AIS', TA'\ >$, TA is the current assessment of the AI system, and AIS' is the updated version of the AI system after incorporating em_z with a new assessment TA'.

Based on the above, TOP aims to find a set of ems that impact positively the TA of an AIS considering a set of TCs:

$$TOP(AIS,\ TCs,\ am,\ em,\ TA) =< \left(em,\ AIS',\ TA'\right)_1,\ \left(em,\ AIS',\ TA'\right)_2,\ \ldots \left(em,\ AIS',\ TA'\right)_n > \tag{6}$$

where n is the number of acceptable em and $TA\prime_n \geq TA\ \forall n$.

4.2. Identify

In the first stage of the process, Identify, all the necessary information is collected and stored in a structured format in the form of cards. The information concerns the overall system (AIS) and its environment, targeting its technical, societal, and business details. To effectively gather and record the required information, all possible AI actors $\left(H^{AIS}\right)$, such as domain experts, technicians, and users, must be involved and provide their inputs where necessary. This stage can be broken down into five sub-processes: (i) System contextualisation, (ii) Information gathering, (iii) Metric and method linking, (iv) Risk and vulnerability deriving, and (v) Documentation information.

4.2.1. System Contextualisation

The system and its environment must be specified comprehensively to provide the context needed for the process to assess and optimise its trustworthiness. The first step is to identify all the AI actors who need to provide input, spanning from business and legal

associates to technicians and domain experts. Their expertise in multiple disciplines is essential to capturing all aspects of the AI system that are related to its trustworthiness. These actors will be tasked with providing input based on their specialty around the following non-exhaustive examples of resources.

- Organisational codes of conduct, guidelines, rules, and procedures;
- Legal requirements and compliance with regulatory frameworks;
- Business environment targets and Key Performance Indicators (KPIs);
- Technical documentation of assets $\left(A^{AIS}\right)$;
- Purpose and specifications of the AI system;
- End-user requirements;
- Possible risks, limitations, and misuse scenarios;
- Third-party agreements, collaborators, and artefacts;
- Environmental and climate concerns;
- User preferences regarding trustworthiness.

4.2.2. Information Gathering

Once the landscape for the system's environment is set, more detailed information must be gathered. The various AI actors are tasked with providing all the necessary information about the system and its assets, including taking into account the previously identified resources. The emphasis of this subprocess is to gather detailed and technical information about the AI system, including the data sources and models that are employed, accompanied by their specifications and relationships with other assets. For a detailed mapping of AI actors and their tasks in the AI system, we refer readers to Appendix A: Descriptions of AI Actor Task (https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A, accessed on 28 February 2025) in the NIST AI RMF knowledge base.

4.2.3. Metric and Method Linking

The different components, processes, and assets of the AI system must be associated with relevant metrics and mitigation methods. This is essential for establishing the *am* and *em* sets for the TCs of interest, which are utilised in subsequent stages of TOP. The former set ensures proper performance monitoring across the TCs, while the latter creates a repository of potential solutions. The system's indicators of focus are mapped to its assets and processes, derived from the use case requirements and KPIs. From there, potential *ams* that produce relevant metrics are identified. Depending on the context, the *ams* can incorporate technical or procedural methods that generate quantitative outputs related to the system's performance and TCs. Furthermore, *ems* are also identified and reviewed to address the limitations, driven by the system's indicators of focus, the metrics produced by the *ams*, and the associated assets. This creates a pathway from the high-level and business-focus specifications to quantitative measurements related to TCs and, subsequently, to practical methods that can be implemented to achieve desired outcomes. For instance, an AI system might be required by legislation or organisational policies to have fair prediction with respect to some sensitive attributes (e.g., age or sex). To efficiently assess or enhance the system based on this requirement, appropriate methods must be identified and employed to its assets. Having an updatable repository of metrics and methods enables the adaptability of the process based on the assets of the AI system. For example, if the AI system has an AI model based on a deep neural network architecture and specific TCs of interest, related *ams* and *ems* that can be employed will be identified and used during the process.

4.2.4. Risk and Vulnerability Deriving

The potential risks are identified based on the expertise of AI actors in both the business and technical aspects of the AI system. Potential vulnerabilities concerning the trustworthiness of the system are characterised in the form of risks. These risks represent high-level concerns that are integrated into the risk management framework employed later in the process. For instance, a business risk may involve user dissatisfaction stemming from discriminatory outcomes of the system. Meanwhile, a technical risk might encompass the leakage of sensitive information due to perturbations in the input data of a model. To ensure consistency and appropriate application, the derived risks must align with the chosen risk management framework. This can result in a knowledge base compatible with the risk management framework and established standards [72–74]. The knowledge base specifies the AI system's components and their interrelations, which are later used to create a model of the system.

4.2.5. Documenting Information

The system's contextual information must be systematically recorded in a structured manner and be accessible during the process. For this reason, the "as-a-card" approach is adopted by TOP to facilitate the documentation of the AI system. In our approach, we consider four main categories of cards that can be employed: use case, data, model, and method cards. In each category, practitioners can incorporate their own set of cards to capture all the information about the system. Even though this subprocess is presented last, the card recording can be conducted gradually and in parallel with the rest of the subprocesses in this stage.

Use case cards. The non-technical, high-level information of the AI system is stored in a use case card. Stakeholders across different scopes of the system are called to record information about the organisation and business perspective such as the purpose, functionality, requirements, objectives, risks, and limitations of the overall AI system. Collected information from this card plays a pivotal role in matching societal and business considerations to technical and quantitative metrics. In addition, it contains information about the involved AI actors and references to other assets of the system that are described from other cards.

Data cards. Technical information about the data is stored in a data card. Primary technical and domain expert stakeholders are called to gather information about the data used inside the AI system. Each dataset has its respective card, highlighting information about the collection, storing, and processing procedures, as well as data samples, including feature descriptions, characteristics, and potential limitations.

Model cards. Technical information about the AI models is stored in a model card. Information regarding the lifecycle of each model must be recorded on this card. This includes the type of the model, hardware requirements, model phases (training, validating, testing), metrics considered, desired outputs, and potential known shortcomings.

Method cards. A method card includes technical and algorithmic information about trustworthiness methods. For each dimension of trustworthiness (fairness, robustness, etc.) individual cards are created that host multiple methods of that category. The information provides a comprehensive description of each method, such as an overall description, its potential limitations and conflicts, AI system lifecycle attributes, requirements, code, and metrics.

Upon completion, the cards provide all the necessary information that is needed to assess the AI system and then find suitable mitigation solutions from applicable methods. In addition, despite having some prerequisite static fields, the cards could be expanded to incorporate specific information depending on the system's particularities or even connect to other components that can actively provide the related information. The selected use

case, data, and model cards must provide information about the stages (S), assets (A), AI actors (H), risks (R), and assessment methods (am) of the system, while the method cards provide the enhancement methods (em). By the completion of the Identify stage, the information is collected, the potential risks are identified, and the cards are filled with the information that defines the system (AIF).

4.3. Assess

After all the needed information is collected, the process moves on to the AI system assessment. At the Assess stage, a quantitative assessment in terms of metrics is executed, which is provided as input for the system's risk assessment. The identified vulnerabilities, technical aspects, and legal/ethical requirements are considered in this process. This stage can be split into the subprocesses of (i) Quantitative Assessment, (ii) Risk Assessment, and iii) Trustworthiness Assessment and Dissemination.

Quantitative Assessment (T)**.** Using the assessment methods (am) from the documentation cards, measurements for the technical components (e.g., datasets and models) can be calculated. The generated measurements need to be linked to the various layers of algorithms, AI models, datasets, and processes of the AI system and provide a clear overview of its technical status. In addition, the assessment of the AI system regarding its legal and ethical requirements takes place in this step. Information from the use case cards is used to classify the system under the legal and ethical regulations and, in conjunction with the metrics, examine its compliance with the related requirements.

Risk Assessment (RM). Upon receipt of the socio-technical context from the cards and the metrics for the respective components, a model of the AI system should be constructed and used by a risk management framework to derive the related risk for trustworthiness. The employed framework should calculate the likelihood and risk level of the threats and consequences for the different TCs (e.g., fairness, robustness) that need to be preserved, and recommend controls to preserve them. The vulnerabilities of the system are closely related to the components' metrics and the controls are mapped to *ems* that mitigate trustworthiness issues.

It is important to note that external and case-specific components can be employed to assess the system. The approaches that will be adopted to generate the quantitative results may vary from only the calculation of metrics and different risk management methods to comprehensive tests from government bodies and regulations.

Trustworthiness Assessment (TA) **and Dissemination.** Upon completion of this stage, the TA of the system is calculated based on the technical and risk assessments, which can be summarised in a report highlighting limitations. After this stage, the AI system stakeholders should be quantitatively aware of the system's shortcomings through the previously identified risks and ready to start improving the overall trustworthiness of the system.

4.4. Explore

Conflicting TCs can pose challenges in finding optimal solutions. Tackling this complicated problem is the main task of the Explore stage, which contributes to the risk treatment step of the risk management process. Utilising the trustworthiness *ems* from the cards, acceptable solutions are searched via exploration. While taking into account the system requirements and assessment metrics, possible sets of solutions are composed and tested iteratively in order to evaluate their performance and impact on the AI system. At the end of this stage, acceptable solution sets that satisfy all the requirements and mitigate as best as possible the related risks are proposed for human approval. The execution of this stage can be divided into (i) Solution Exploration and (ii) Solution Recommendation.

Solution Exploration. Based on the current stage of the AI system lifecycle and its characteristics, applicable methods can be found by filtering the method cards. Depending on the system specificities, solutions are generated including a single or multiple *ems* to be applied to the AI system. These solutions must be tested in a development environment and undergo the assessment procedure of the previous stage of TOP to calculate their *TA*. Finding a solution for the trustworthiness problem poses difficulties and limitations, but the exploration mechanism of the stage can be substituted with different approaches depending on the current status. For instance, if the pool of methods is shallow, one can exhaustively test all the different combinations and find the best possible solution. In contrast, if the brute force approach is not viable, a heuristic or predictive approach can also be employed to find optimal solutions effectively.

Solution Recommendation. Testing multiple solutions for mitigating trustworthiness issues can yield a diverse set of resulting metrics and risk levels that need to be considered before selecting the best solution. Initially, some solutions might be excluded if they compromise any of the requirements or objectives of the system. Comparing the remaining acceptable solutions is non-trivial due to their different impact and potential conflicts on the TCs. To resolve this issue, multi-criteria decision-making (MCDM) methods (e.g., TOPSIS and VIKOR) can be employed to assist with the decision of which solutions best meet the current requirements [75]. To transform this into a multi-criteria problem, a decision matrix is necessary, based on the tested solutions from the *ems* and their indicators from the *TA*. To construct the matrix, the following formulation can be followed for a set of m solutions and n indicators.

- The solutions sets are represented by the alternatives A_j where $i = 1, 2, \ldots, m$;
- The indicators, such as metrics and risk levels from TA, are represented by the criteria C_j where $j = 1, 2, \ldots n$;
- The organisational or supervisor preferences from the actors H of the AIS are represented by the weights w_j, associated with each C_j and $\sum_{j=1}^{n} w_j = 1$;
- The performance of each alternative A_j, with respect to the criterion C_j, is represented by x_{ij}.

Based on the above, the following decision matrix X is defined:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

After the decision matrix is constructed, any applicable MCDM method can be employed to evaluate and rank the solution sets and recommend which solution fits best the system's needs.

4.5. Enhance

Moving on to the final stage, human supervisors are informed about the current status of the AI system from the Assess stage and what methods (ems) can be employed, accompanied by their trustworthiness assessment (AT) from the Explore stage. Based on these, they are called to select which solution will be applied to the system. The Enhance stage can be divided into two subprocess: (i) Applying Solutions and (ii) Update and Monitor.

Applying Solutions. After a proposed solution is approved, it is applied while making all the necessary changes to the AI system (AIS) and the respective documentation cards. Depending on the enhancement methods (ems) that compromise the selected solution, the algorithms are implemented to the respective assets of the system manually or

automatically, and their implications to the system must be reflected in the documentation cards. The Enhance stage, however, is not necessarily the end of the process because TOP is a continuous process that assesses and enhances the system's trustworthiness.

Update and Monitor. In this stage, the updated version of the AI system ($AIS\prime$) is continually monitored and the documentation cards are updated based on the specified metrics and risks. The observations are used as input to the Identify stage, which can cause the process to kickstart again in an iterative manner due to performance changes that can compromise system requirements, user requests, and KPIs. Upon completion of the Enhance stage, assuming the solution is satisfactory, the system should be ameliorated in terms of trustworthiness.

Communication and human oversight stand as critical aspects of TOP throughout its execution and during the dissemination of information through reports and monitoring mechanisms. Equally important is ensuring that humans receive interpretable information that highlights key points of attention and includes adequate explanations. The concept of explainability can be examined from two perspectives. The first one focuses on the explainability of the AI system's models, where it is considered as one of the TCs that need to be preserved. To address relevant risks, appropriate *ams* and *ems* must be identified. For instance, approaches such as Shapley Additive exPlanations (SHAP) [76], Local Interpretable Model-agnostic Explanations (LIME) [77], and counternatural explanations [78] can be employed for specific AI models. The second perspective revolves around the implementation of TOP and the overall use case, focusing more on how the information is presented to humans. While this could be addressed by leveraging the explainability of AI models, it is crucial provide sufficient details about the underlying process and outcomes of TOP that serve as explanations. For example, this could involve details about the output of *ams* and their intuitive interpretations, supported by visualisations to facilitate understanding across the impacts between different *ems*.

5. Case Study

To demonstrate the practical application of the proposed TOP, a case study has been conducted utilising a widely used dataset for fairness analysis and mitigation. The analysis includes metrics to assess bias and discrimination in the data and predictions between sensitive variables of the dataset, and how those shortcomings can be mitigated by employing trustworthiness methods. Three hypothetical scenarios have been constructed around the dataset, as presented in Table 1, differentiated by the status of the AI system. In the first scenario (S1), the AI system is at the design stage of the lifecycle; in the second scenario (S2), it is at the development stage; and in the third scenario (S3), it is at the deployment stage. The variance in the lifecycle stage is reflected in the available assets, completed cards, and characteristics of the AI system.

Table 1. Case study scenarios and details.

Use Case	AI System Lifecycle Stage (*S*)	Available Assets (*A*)	Completed Cards	Lifecycle Characteristic
S1—Design	Design	Application Domain, Data	Use case, Data	pre-processing
S2—Develop	Develop	Application Domain, Data, AI Model	Use case, Data, Model	pre-processing, in-processing
S3—Deploy	Deploy	Application Domain, Data, AI Model	Use case, Data, Model	pre-processing, in-processing, post-processing

Case Study on Fairness Assessment: Income Prediction (Adult Dataset). The Adult dataset [79] is used for binary classification tasks. The objective is to predict whether an individual's annual income exceeds USD 50,000 based on personal characteristics, such as

demographic and employment details from the 1994 US Census Bureau. A use case has been constructed where a bank wants to implement a machine learning model to predict clients' annual income. In accordance with new regulations, the bank is required to assess its predictive system for bias risks and mitigate them if present. For the AI system, a simple pipeline has been considered where a Logistic Regression model is trained on the dataset to predict whether an individual's income exceeds USD 50,000 annually.

Metrics and Methods. To evaluate the fairness of the AI system and test mitigation algorithms, the AI Fairness 360 (AIF360) toolkit by IBM [80] has been selected. This toolkit provides concrete implementations of metrics and methods from the literature for assessing and mitigating fairness issues of well-established datasets, such as the Adult dataset. In this case study, the metrics represent the *ams*, and the methods represent the *ems*. Detailed lists of the *am* and *em* considered in this study can be found in Tables 2 and 3, respectively. For the *ams*, the acceptable range has also been included, representing the requirements of the AI system in the case study.

Table 2. Descriptions and acceptable ranges of fairness assessment methods. P and N represent the total number of positive and negative outcomes; I is the number of instances; un and pr denote the calculation for the unprivileged and privileged groups, respectively; True positive (TP): number of positive predictions where the actual outcome is positive; True negative (TN): number of negative predictions where the actual outcome is negative; False positive (FP): number of positive predictions where the actual outcome is negative; False negative (FN): number of negative predictions where the actual outcome is positive; False positive rate (FPR): number of FP divided by the number of actual positives; True positive rate (TPR): number of TP divided by the number of actual positives; m and r denote the calculation for the monitored and reference group, respectively; b_i is the subtraction between the predicted and actual outcome, plus 1 for the individual i; and μ denotes to the mean of the b_i values.

Name	Definitions	Description	Ideal Value	Acceptable Range
BALANCED ACCURACY	$\frac{TP+TN}{P+N}$	Accuracy metric for the classifier	-	>0.7
STATISTICAL PARITY DIFFERENCE	$\left(\frac{P}{I}\right)^{un} - \left(\frac{P}{I}\right)^{pr}$	Difference of the rate of favourable outcomes received by the unprivileged group to the privileged group	0	$[-0.1, 0.1]$
DISPARATE IMPACT	$\frac{\left(\frac{P}{I}\right)^{un}}{\left(\frac{P}{I}\right)^{pr}}$	Ration of rate of favourable outcome for the unprivileged group to that of the privileged group.	1	[0.7, 1.3]
AVERAGE ODDS DIFFERENCE	$\frac{(FPR^m - FPR^r) + (TPR^m - TPR^r)}{2}$	Average difference of false positive rate and true positive rate between unprivileged and privileged groups	0	$[-0.1, 0.1]$
EQUAL OPPORTUNITY DIFFERENCE	$\lvert TPR^{pr} - TPR^{un} \rvert$	Difference of true positive rates between the unprivileged and privileged groups	0	$[-0.1, 0.1]$
THEIL INDEX	$\frac{1}{n}\sum_{i=1}^{n} \frac{b_i}{\mu} ln \frac{b_i}{\mu}$	Generalised entropy of benefit for all individuals in the dataset; measures the inequality in benefit allocation for individuals	0	-

Table 3. Enhancement methods related to fairness.

Pre-Processing	In-Processing	Post-Processing
Optimised pre-processing [81]	Adversarial debaising [8]	Calibrated equalized odds [82]
Reweighting [7]	Prejudice remover [83]	Rejection option classification [84]
Learning fair representations [85]	Gerry-Fair (FairFictPlay) [86]	
Disparate impact remover [87]		

5.1. Identify

In this stage of the process, information is collected and documented on the respective cards. Multiple AI actors (H), including the board members, managers, bankers, technical employees, legal advisors, and users of the AI system, contribute to their completion. The recorded information is guided by the system's objectives and encompasses the requirements it must fulfil, as well as relevant metrics to be measured. For instance, certain requirements might revolve around the accuracy, response time, availability, and fairness of the AI system, which can be validated using the specified metrics. Table 4 provides the card status for each scenario.

Table 4. Card status of case study scenarios.

Scenarios	Cards			
	Use Case	**Data**	**Model**	**Method**
S1—Design	Partially	Partially	Incomplete	Partially
S2—Develop	Partially	Completed	Partially	Completed
S3—Deploy	Completed	Completed	Completed	Completed

S1-Design. In this scenario, only the domain application and data assets are available. Consequently, the use case, data, and method cards for the AI system are completed. Within these cards, the respective stakeholders document the intended usage of the system, potential risks (R), and preliminary information about the data used. Given the preliminary status of the AI system, the cards may be partially completed due to the unavailability of certain information. For instance, the absence of an AI model restricts the recordable *ems* to a subset that focuses solely on the data aspect.

S2-Develop and S3-Deploy. Both scenarios are similar in the Identify stage, with the assets of domain application, data, and AI model being available, resulting in the completion of all the cards. In S2, the basic technical information about the AI model is expected, while in S3 additional information regarding its deployment is recorded. In each scenario, the use case and data cards are updated to reflect any changes to the purpose and risks of the AI system based on the adopted AI model and pipeline. In addition, the method card is populated with additional *ems* in accordance with the progressive characteristics of the AI system.

5.2. Assess

In this stage, the quantitative and risk assessment of the dataset and AI model is performed. The selected *am* are derived from the information documented in the cards, including the objectives and risks (R) of the AI system, as well as the types of data and models employed. By integrating the generated quantitative measurements with a risk management framework, the risks on trustworthiness can be calculated. For this case study, the risk of the AI system is considered to be biased in terms of the sensitive variable of sex. In a more complex use case, a comprehensive risk management framework can be

employed to calculate a variety of risks for the AI system. The quantitative assessment results for each scenario can be found in Table 5.

Table 5. Case study quantitative assessment results.

Scenarios	Metrics					
	Accuracy	Statistical Parity Difference	Disparate Impact	Average Odds Difference	Equal Opportunity Difference	THEIL Index
S1—Design	-	−0.1532	0.5949	-	-	-
S2—Develop, S3—Deploy	0.7437	−0.3580	0.2794	−0.3181	−0.3768	0.1129

S1-Design. In this scenario, due to the absence of the AI model, applicable *am* are limited. From the dataset, differences between the privileged and unprivileged groups, with respect to their sensitive variables, are calculated. Specifically, the statistical parity difference and the disparate impact are calculated between the groups.

S2-Develop and S3-Deploy. In these scenarios, an AI model is trained enabling the usage of additional *am* that focus on the final predictions. These *am* include the Average Odds Difference, Equal Opportunity Difference, and the THEIL index.

After calculating the metrics and risks, a comprehensive report can be generated, providing various stakeholders with the status of the AI system's trustworthiness. Descriptions and explanations can accompany the report to inform stakeholders from different backgrounds.

5.3. Explore

In this step, the main distinction between the three scenarios is observed. Each scenario considers different characteristics of the AI system, thereby different *em* are applicable. By querying the documented algorithms from the method cards with the characteristics of the system, as presented in Table 1, an exploration of applicable *em* can be conducted. For each scenario, the identified *ems* are tested to observe their effect on the AI system by repeating the assessment performed in the previous stage. In this case study, the technical assessment (T) of the system is used to construct the decision matrix where the indicators are represented by the metrics produced from the *ams*, and the alternatives are represented by the *ems*.

S1-Design. As previously stated, this scenario is limited to the data asset and the pre-processing *ems* for mitigating trustworthiness issues. The results obtained after executing the applicable *ems* can be observed in Table 6. Based on the desired values for the metrics, the clear choice in this instance is the Reweighting method, as it achieves optimal values for both metrics.

Table 6. Scenario 1—Design method metrics. Bolded numbers highlight the best performing metrics.

Method	Metrics	
	Statistical Parity Difference	Disparate Impact
No method	−0.1902	0.3677
Reweighting	**0.0**	**1.0**
Disparate impact remover	−0.1962	0.3580
Optimized pre-processing	−0.0473	0.8199

S2-Develop and S3-Deploy. In these scenarios, the entire pipeline of the AI system can be evaluated, producing multiple metrics for each *em* used. The sole distinction between the two scenarios is the inclusion of the last two *em* (calibrated equalized odds

and rejection option classification), which take place only in S3-Deploy. The metric values can be observed in Table 7. The denoted NaN values for the Disparate Impact metric occur when the denominator equals zero, due to the absence of positive prediction outcomes for a group characterised by the sensitive variables.

Table 7. Scenario 2—Develop and Scenario 3—Deploy assessment metrics. Bolded numbers highlight the best performing metrics.

Method	Scenarios Applicability	Metrics					
		Accuracy	Statistical Parity Difference	Disparate Impact	Average Odds Difference	Equal Opportunity Difference	THEIL Index
Reweighting	S2—Develop, S3—Deploy	0.7133	−0.0705	0.7785	0.0188	0.0293	0.1400
Optimal pre-processing	S2—Develop, S3—Deploy	0.7153	−0.0962	0.7207	**−0.0119**	**−0.0082**	0.1366
Disparate impact remover	S2—Develop, S3—Deploy	**0.7258**	**−0.0382**	0.8965	0.0559	0.0639	0.1224
Adversarial debiasing	S2—Develop, S3—Deploy	0.6637	−0.2095	0.0	−0.279	−0.4595	0.1793
Prejudice remover	S2—Develop, S3—Deploy	0.6675	−0.2184	0.0	−0.2900	−0.4734	0.1769
Gerry-Fair (FairFictPlay)	S2—Develop, S3—Deploy	0.4698	0.0	NaN	0.0	0.0	0.2783
Calibrated equalized odds	S3—Deploy	0.5	0.0	NaN	0.0	0.0	0.2783
Rejection option classification	S3—Deploy	0.7140	−0.0402	**0.9088**	0.0423	0.0407	**0.1171**

In both cases, identifying the best *em* to implement is non-trivial. Methods may excel in certain metrics while falling short in others. The metric differences between *ems* can be also observed in an illustrative manner in the radar plot of Figure 5. To address this issue, *ems* that do not meet the requirements of the AI system are excluded and a MCDM method is employed to facilitate the selection from the remaining *ems*. The methods of Adversarial debiasing, Prejudice remover, Gerry-Fair, and Calibrated odds are excluded because they compromise the acceptable range for accuracy, and some of them negatively affect the metrics of Average Odds Difference and Equal Opportunity Difference.

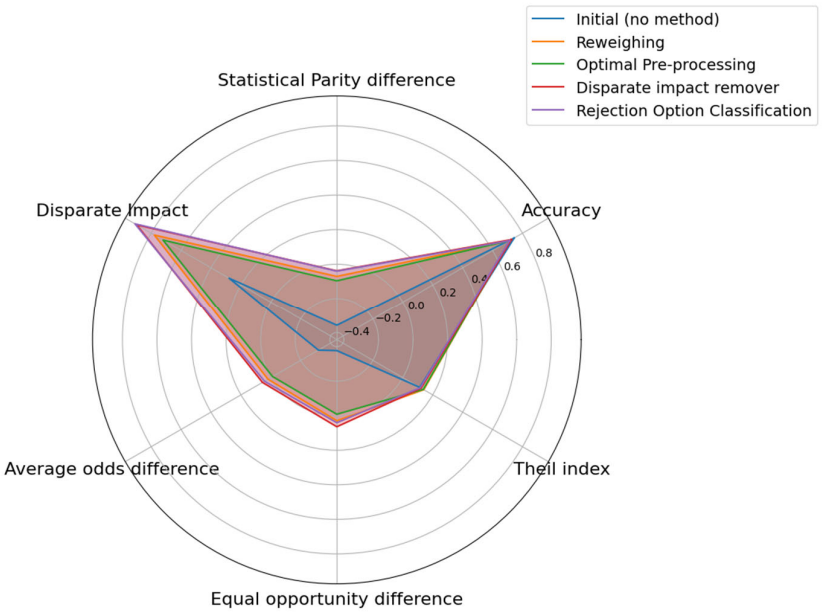

Figure 5. Radar plot of method decision matrix.

Subsequently, an MCDM is utilised by constructing the decision matrix from the remaining options. In the matrix, the *ems* and the initial setup serve as the alternatives, while the metrics function as criteria. Moreover, a weight array is constructed to represent the preferences of the different stakeholders. For the experiments, the VIKOR method was used with the parameter v = 0.5.

As observed from the results in Table 8, the best method to select can vary depending on the preferences of the metrics. This indicates that through the weights, the metrics can be tailored to meet the specific needs of the AI system, and that both the selected metrics and the preferences of the stakeholders can significantly impact the ultimate method selected for implementation. Moreover, the choice of the MCDM and its parameters plays a crucial role, as different configurations may yield different results. In any case, the outlined approach provides stakeholders with essential information for deciding which methods can be used to enhance the trustworthiness of the AI system.

Table 8. Recommended solutions from MCDM.

Scenario	Weights	Selected *em*
S2—Develop	[16.6, 16.6, 16.6, 16.6, 16.6, 16.6]	Disparate impact remover
	[80, 4, 4, 4, 4, 4]	Initial
	[4, 4, 4, 80, 4, 4]	Optimal pre-processing, Disparate impact remover
S3—Deploy	[16.6, 16.6, 16.6, 16.6, 16.6, 16.6]	Disparate impact remover
	[4, 4, 4, 80, 4, 4]	Optimal pre-processing, Disparate impact remover

5.4. Enhance

Upon deciding which *em* will be employed to mitigate the trustworthiness risk of the AI system, it must be integrated into the pipeline, and all relevant cards must be updated to reflect the current status of the system. If a pre-processing method that alters a dataset was introduced, the respective data card must be updated to reflect this change. The same applies for the model and use case cards, where the updated pipeline can result in changes to specific documented information such as the model metrics and the requirements of the AI system. Implementing the method em_z will produce an updated version of the system $AIS\prime$ and its assessment $TA\prime$ where:

$$TOP(AIS,\ TCs,\ am,\ em,\ TA) \rightarrow em_z\Big(TCs_f, AIS,\ TA\Big) = AIS^{'}, TA^{'}$$

Any modifications made must be propagated through the subsequent stages of the lifecycle, where the proposed process will be re-executed until the Monitor stage of the lifecycle. At that stage, the AI system must be continually monitored and assessed to ensure its performance does not degrade. If any anomalies or deviations from its intended purpose and behaviour are observed, the AI system must undergo the proposed process again to mitigate any identified risks.

6. Discussion and Conclusions

In this work, we proposed TOP, a comprehensive methodology for assessing and enhancing the trustworthiness of AI systems that augments their lifecycle with the perspective of TAI. Similar procedural methodologies vary on their level of granularity and do not consider some of the key concepts in the field of TAI. The capAI methodology considers a structured lifecycle-wide conformity assessment. Z-Inspection® takes a holistic socio-technical approach to evaluate trustworthiness and provides recommendations to

stakeholders. The AI Assessment Catalog offers guidelines evaluation thought a stepwise risk management process. ECCOLA focuses on the integration between the AI development and ethical awareness, but lacks a formal risk management assessment. POLARIS consolidates best practices for the development of the AI system, but it is still in preliminary stages. Our approach differentiates by combining the concepts of documentation cards and risk management into a unified process that augments the whole AI system lifecycle in order to practically assess and enhance its trustworthiness. The process operationalises efforts for making AI systems trustworthy through the utilisation of procedural methodologies and algorithms (metrics and methods) and the involvement of relevant AI actors. We demonstrated how documentation cards can be employed to capture essential information and facilitate the execution of TOP. Through three scenarios in a case study, we addressed the lifecycle dimension challenge by demonstrating the applicability and adaptability of the process through the lifecycle of an AI system by improving its trustworthiness using its characteristics and assets. Furthermore, the use of cards to capture a broad range of socio-technical information and the application of MCDM methods contribute to conflict management among the TCs. Additionally, the inclusion and augmentation of AI actor decision-making within the process promotes human oversight and collaboration.

Our experimental results validate the functionality of the process; however, certain limitations related to its practical applicability must be acknowledged. While the case study demonstrates TOP's potential, the experiments were conducted using hypothetical AI systems that considered only one TC and did not incorporate a comprehensive risk management framework. A methodology for achieving TAI must be universally applicable across all types of AI models and systems. Although feedback from three distinct use cases was incorporated during TOP's development, its generalizability may still be limited in certain scenarios. Another limitation is associated with the implementation of trustworthiness in AI system and the overhead that it could cause in terms of increased computation and execution time costs. For instance, running algorithms for the assessment and enhancement of the AI system can require extensive computational resources, potentially leading to response delays of the system in time critical scenarios. While these challenges are not necessarily inherent limitations of TOP itself, shortcomings could arise due to its lack of scalability and automation. Ensuring that the approach can be applied efficiently in large-scale systems and automating its procedures responsibly can result in unforeseen challenges. As mentioned, the conflicts between the TCs and how to resolve them are still unresolved challenges in the field of TAI highly impacted by the current SOTA techniques and the AI system's environment. This issue has attracted the attention of researchers who try to catalogue [88] and resolve [89] these conflicts and tensions between the TCs. Our methodology employs an MCDM approach to navigate conflicts and help humans make informed decisions about the enhancement of the AI system, but it does not comprehensively resolve the trade-offs between the TCs. New novel approaches from the literature could be leveraged to augment or even replace the MCDM approach of TOP to effectively resolve and manage the trade-offs.

To address these limitations, future work will focus on applying TOP in real-world systems, considering multiple TCs and incorporating a concrete risk management framework tailored to trustworthiness. Specifically, we intend to apply TOP in three distinct use cases: a (i) maritime port, (ii) media institution, and (iii) medical hospital. This will enable the identification and resolution of potential shortcomings of TOP. Finally, we believe that the emerging capabilities of AI that cause mistrust between humans and AI can also be part of the solution. For this reason, our future work will also focus on leveraging the advantages of symbolic [90] and agentic [91] approaches to further augment TOP and the support towards decision-making in the context of TAI. Incorporating tailored AI agents to TOP

can potentially resolve challenges related to the explainability, scalability, and automation of the proposed method.

Author Contributions: Conceptualization, M.F., K.L., D.A. and G.M.; methodology, M.F. and K.L.; investigation, M.F.; software, M.F.; supervision, D.A. and G.M.; writing—original draft, M.F.; writing—review and editing, K.L., D.A. and G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the EU Horizon Europe programme CL4-2022-HUMAN-02-01 under the project THEMIS 5.0 (grant agreement No. 101121042) and by the UK Research and innovation under the UK governments Horizon funding guarantee. The work presented here reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

Data Availability Statement: The original data presented in the study are openly available in the "Adult" repository at https://archive.ics.uci.edu/dataset/2/adultor (accessed on 26 February 2025).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Singla, A.; Sukharevsky, A.; Yee, L.; Chui, M.; Hall, B. *The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value*; McKinsey and Company: Brussels, Belgium, 2024.
2. Baeza-Yates, R.; Fayyad, U.M. Responsible AI: An Urgent Mandate. *IEEE Intell. Syst.* **2024**, *39*, 12–17. [CrossRef]
3. Mariani, R.; Rossi, F.; Cucchiara, R.; Pavone, M.; Simkin, B.; Koene, A.; Papenbrock, J. Trustworthy AI—Part 1. *Computer* **2023**, *56*, 14–18. [CrossRef]
4. Díaz-Rodríguez, N.; Del Ser, J.; Coeckelbergh, M.; de Prado, M.L.; Herrera-Viedma, E.; Herrera, F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf. Fusion* **2023**, *99*, 101896. [CrossRef]
5. Prem, E. From Ethical AI Frameworks to Tools: A Review of Approaches. *AI Ethics* **2023**, *3*, 699–716. [CrossRef]
6. Narayanan, M.; Schoeberl, C. A Matrix for Selecting Responsible AI Frameworks. Center for Security and Emerging Technology. 2023. Available online: https://cset.georgetown.edu/publication/a-matrix-for-selecting-responsible-ai-frameworks/#:~:text=The%20matrix%20provides%20a%20structured,to%20guidance%20that%20already%20exists (accessed on 26 February 2025).
7. Kamiran, F.; Calders, T. Data Preprocessing Techniques for Classification without Discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
8. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 1–3 February 2018; pp. 335–340. [CrossRef]
9. Hardt, M.; Price, E.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2016; Volume 29. Available online: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html (accessed on 26 February 2025).
10. Boza, P.; Evgeniou, T. *Implementing Ai Principles: Frameworks, Processes, and Tools*; SSRN Scholarly Paper; SSRN: Rochester, NY, USA, 2021. [CrossRef]
11. Floridi, L.; Holweg, M.; Taddeo, M.; Amaya, J.; Mökander, J.; Wen, Y. *capAI—A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*; SSRN Scholarly Paper; SSRN: Rochester, NY, USA, 2022. [CrossRef]
12. Mökander, J.; Schuett, J.; Kirk, H.R.; Floridi, L. Auditing Large Language Models: A Three-Layered Approach. *AI Ethics* **2024**, *4*, 1085–1115. [CrossRef]
13. Mökander, J.; Floridi, L. Operationalising AI governance through ethics-based auditing: An industry case study. *AI Ethics* **2022**, *3*, 451–468. [CrossRef]
14. Zicari, R.V.; Brodersen, J.; Brusseau, J.; Dudder, B.; Eichhorn, T.; Ivanov, T.; Kararigas, G.; Kringen, P.; McCullough, M.; Moslein, F.; et al. Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Trans. Technol. Soc.* **2021**, *2*, 83–97. [CrossRef]
15. Allahabadi, H.; Amann, J.; Balot, I.; Beretta, A.; Binkley, C.; Bozenhard, J.; Bruneault, F.; Brusseau, J.; Candemir, S.; Cappellini, L.A.; et al. Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients. *IEEE Trans. Technol. Soc.* **2022**, *3*, 272–289. [CrossRef]
16. Poretschkin, M.; Schmitz, A.; Akila, M.; Adilova, L.; Becker, D.; Cremers, A.B.; Hecker, D.; Houben, S.; Mock, M.; Rosenzweig, J.; et al. Guideline for Trustworthy Artificial Intelligence—AI Assessment Catalog. *arXiv* **2023**, arXiv:2307.03681.

17. Vakkuri, V.; Kemell, K.-K.; Jantunen, M.; Halme, E.; Abrahamsson, P. ECCOLA—A method for implementing ethically aligned AI systems. *J. Syst. Softw.* **2021**, *182*, 111067. [CrossRef]
18. Nasr-Azadani, M.M.; Chatelain, J.-L. The Journey to Trustworthy AI- Part 1: Pursuit of Pragmatic Frameworks. *arXiv* **2024**, arXiv:2403.15457.
19. Brunner, S.; Frischknecht-Gruber, C.M.-L.; Reif, M.; Weng, J. A Comprehensive Framework for Ensuring the Trustworthiness of AI Systems. In *Proceeding of the 33rd European Safety and Reliability Conference*; Research Publishing Services: Southampton, UK, 2023; pp. 2772–2779. [CrossRef]
20. Baker-Brunnbauer, J. TAII Framework. In *Trustworthy Artificial Intelligence Implementation: Introduction to the TAII Framework*; Springer International Publishing: Cham, Switzerland, 2022; pp. 97–127.
21. Guerrero Peño, E. *How the TAII Framework Could Influence the Amazon's Astro Home Robot Development*; SSRN Scholarly Paper; SSRN: Rochester, NY, USA, 2022. [CrossRef]
22. Baldassarre, M.T.; Gigante, D.; Kalinowski, M.; Ragone, A. POLARIS: A Framework to Guide the Development of Trustworthy AI Systems. In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering—Software Engineering for AI, CAIN'24, Lisbon, Portugal, 14–15 April 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 200–210. [CrossRef]
23. Stettinger, G.; Weissensteiner, P.; Khastgir, S. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. *IEEE Access* **2024**, *12*, 22718–22745. [CrossRef]
24. Confalonieri, R.; Alonso-Moral, J.M. An Operational Framework for Guiding Human Evaluation in Explainable and Trustworthy Artificial Intelligence. *IEEE Intell. Syst.* **2023**, *39*, 18–28. [CrossRef]
25. Hohma, E.; Lütge, C. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI* **2023**, *4*, 904–925. [CrossRef]
26. Ronanki, K.; Cabrero-Daniel, B.; Horkoff, J.; Berger, C. RE-Centric Recommendations for the Development of Trustworthy(Er) Autonomous Systems. In Proceedings of the First International Symposium on Trustworthy Autonomous Systems, TAS '23, Edinburgh, UK, 11–12 July 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–8. [CrossRef]
27. Microsoft. Responsible AI Transparency Report: How We Build, Support Our Customers, and Grow. 2024. Available online: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2024.pdf (accessed on 26 February 2025).
28. PwC. From Principles to Practice: Responsible AI in Action. 2024. Available online: https://www.strategy-business.com/article/From-principles-to-practice-Responsible-AI-in-action (accessed on 26 February 2025).
29. Digital Catapult. Operationalising Ethics in AI. 20 March 2024. Available online: https://www.digicatapult.org.uk/blogs/post/operationalising-ethics-in-ai/ (accessed on 26 February 2025).
30. IBM. Scale Trusted AI with Watsonx. Governance. 2024. Available online: https://www.ibm.com/products/watsonx-governance (accessed on 26 February 2025).
31. Accenture. Responsible AI: From Principles to Practice. 2021. Available online: https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice (accessed on 26 February 2025).
32. Vella, H. Accenture, AWS Launch Tool to Aid Responsible AI Adoption. *AI Business*, 28 August 2024. Available online: https://aibusiness.com/responsible-ai/accenture-aws-launch-tool-to-aid-responsible-ai-adoption (accessed on 26 February 2025).
33. Micheli, M.; Hupont, I.; Delipetrev, B.; Soler-Garrido, J. The landscape of data and AI documentation approaches in the European policy context. *Ethic-Inf. Technol.* **2023**, *25*, 56. [CrossRef]
34. Königstorfer, F. A Comprehensive Review of Techniques for Documenting Artificial Intelligence. *Digit. Policy Regul. Gov.* **2024**, *26*, 545–559. [CrossRef]
35. Oreamuno, E.L.; Khan, R.F.; Bangash, A.A.; Stinson, C.; Adams, B. The State of Documentation Practices of Third-Party Machine Learning Models and Datasets. *IEEE Softw.* **2024**, *41*, 52–59. [CrossRef]
36. Pushkarna, M.; Zaldivar, A.; Kjartansson, O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 1776–1826. [CrossRef]
37. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; ACM: Atlanta, GA, USA, 2019; pp. 220–229. [CrossRef]
38. Crisan, A.; Drouhard, M.; Vig, J.; Rajani, N. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Seoul, Republic of Korea, 21–24 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 427–439. [CrossRef]
39. Liang, W.; Rajani, N.; Yang, X.; Ozoani, E.; Wu, E.; Chen, Y.; Smith, D.S.; Zou, J. What's Documented in AI? Systematic Analysis of 32K AI Model Cards. *arXiv* **2024**, arXiv:2402.05160.

40. Alsallakh, B.; Cheema, A.; Procope, C.; Adkins, D.; McReynolds, E.; Wang, E.; Pehl, G.; Green, N.; Zvyagina, P. System-Level Transparency of Machine Learning. Meta AI. 2022. Available online: https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/ (accessed on 26 February 2025).
41. Meta. Introducing 22 System Cards That Explain How AI Powers Experiences on Facebook and Instagram. 29 June 2023. Available online: https://ai.meta.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/ (accessed on 26 February 2025).
42. Pershan, C.; Vasse'i, R.M.; McCrosky, J. This Is Not a System Card: Scrutinising Meta's Transparency Announcements. 1 August 2023. Available online: https://foundation.mozilla.org/en/blog/this-is-not-a-system-card-scrutinising-metas-transparency-announcements/ (accessed on 26 February 2025).
43. Hupont, I.; Fernández-Llorca, D.; Baldassarri, S.; Gómez, E. Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act. *arXiv* **2023**, arXiv:2306.13701. [CrossRef]
44. Gursoy, F.; Kakadiaris, I.A. System Cards for AI-Based Decision-Making for Public Policy. *arXiv* **2022**, arXiv:2203.04754.
45. Golpayegani, D.; Hupont, I.; Panigutti, C.; Pandit, H.J.; Schade, S.; O'Sullivan, D.; Lewis, D. AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. In *Privacy Technologies and Policy*; Jensen, M., Lauradoux, C., Rannenberg, K., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2024; Volume 14831, pp. 48–72. [CrossRef]
46. Giner-Miguelez, J.; Gómez, A.; Cabot, J. Using Large Language Models to Enrich the Documentation of Datasets for Machine Learning. *arXiv* **2024**, arXiv:2404.15320.
47. Yang, X.; Liang, W.; Zou, J. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. *arXiv* **2024**, arXiv:2401.13822.
48. Mehta, S.; Rogers, A.; Gilbert, T.K. Dynamic Documentation for AI Systems. *arXiv* **2023**, arXiv:2303.10854.
49. Chmielinski, K.; Newman, S.; Kranzinger, C.N.; Hind, M.; Vaughan, J.W.; Mitchell, M.; Stoyanovich, J.; McMillan-Major, A.; McReynolds, E.; Esfahany, K. The CLeAR Documentation Framework for AI Transparency. Harvard Kennedy School Shorenstein Center Discussion Paper 2024. Available online: https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/ (accessed on 26 February 2025).
50. Wing, J.M. Trustworthy AI. *Commun. ACM* **2021**, *64*, 64–71. [CrossRef]
51. AI HLEG. Ethics Guidelines for Trustworthy AI. 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 26 February 2025).
52. ENISA; Adamczyk, M.; Polemi, N.; Praça, I.; Moulinos, K. *A Multilayer Framework for Good Cybersecurity Practices for AI: Security and Resilience for Smart Health Services and Infrastructures*; Adamczyk, M., Moulinos, K., Eds.; European Union Agency for Cybersecurity: Athens, Greece, 2023. [CrossRef]
53. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST: Gaithersburg, MD, USA, 2023.
54. Kaur, D.; Uslu, S.; Rittichier, K.J.; Durresi, A. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]
55. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **2023**, *55*, 1–46. [CrossRef]
56. U.S.G.S.A. *Understanding and Managing the AI Lifecycle. AI Guide for Government*; 2023. Available online: https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/ (accessed on 26 February 2025).
57. OECD. *OECD Framework for the Classification of AI Systems*; OECD Digital Economy Papers, No. 323; OECD: Paris, France, 2022.
58. Calegari, R.; Castañé, G.G.; Milano, M.; O'Sullivan, B. Assessing and Enforcing Fairness in the AI Lifecycle. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macau, China, 19–25 August 2023; International Joint Conferences on Artificial Intelligence Organization: Macau, China, 2023; pp. 6554–6662. [CrossRef]
59. Schlegel, M.; Sattler, K.-U. Management of Machine Learning Lifecycle Artifacts: A Survey. *arXiv* **2022**, arXiv:2210.11831.
60. Toreini, E.; Aitken, M.; Coopamootoo, K.P.L.; Elliott, K.; Zelaya, V.G.; Missier, P.; Ng, M.; van Moorsel, A. Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context. *arXiv* **2022**, arXiv:2007.08911.
61. Suresh, H.; Guttag, J.V. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *arXiv* **2021**, arXiv:1901.10002.
62. Liu, H.; Wang, Y.; Fan, W.; Liu, X.; Li, Y.; Jain, S.; Liu, Y.; Jain, A.K.; Tang, J. Trustworthy AI: A Computational Perspective. ACM Trans. *Intell. Syst. Technol.* **2022**, *14*, 1–59. [CrossRef]
63. Mentzas, G.; Fikardos, M.; Lepenioti, K.; Apostolou, D. Exploring the landscape of trustworthy artificial intelligence: Status and challenges. *Intell. Decis. Technol.* **2024**, *18*, 837–854. [CrossRef]
64. Zhao, W.; Alwidian, S.; Mahmoud, Q.H. Adversarial Training Methods for Deep Learning: A Systematic Review. *Algorithms* **2022**, *15*, 283. [CrossRef]
65. OECD. *Advancing Accountability in AI: Governing and Managing Risks Throughout the Lifecycle for Trustworthy AI*; OECD Digital Economy Papers, No. 349; OECD Publishing: Paris, France, 2023. [CrossRef]

66. EU AI Act: First Regulation on Artificial Intelligence. Topics | European Parliament. 8 June 2023. Available online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 26 February 2025).
67. *ISO/IEC TR 24028:2020(En)*; Information Technology—Artificial Intelligence—Overview of Trustworthiness in Artificial Intelligence. ISO: Geneva, Switzerland, 2020. Available online: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:24028:ed-1:v1:en (accessed on 2 December 2024).
68. Mannion, P.; Heintz, F.; Karimpanal, T.G.; Vamplew, P. Multi-Objective Decision Making for Trustworthy AI. In Proceedings of the Multi-Objective Decision Making (MODeM) Workshop, Online, 14–16 July 2021.
69. Alsalem, M.; Alamoodi, A.; Albahri, O.; Albahri, A.; Martínez, L.; Yera, R.; Duhaim, A.M.; Sharaf, I.M. Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. *Expert Syst. Appl.* **2024**, *246*, 123066. [CrossRef]
70. Mattioli, J.; Sohier, H.; Delaborde, A.; Pedroza, G.; Amokrane-Ferka, K.; Awadid, A.; Chihani, Z.; Khalfaoui, S. Towards a Holistic Approach for AI Trustworthiness Assessment Based upon Aids for Multi-Criteria Aggregation. In Proceedings of the SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety (Vol. 3381), Washington, DC, USA, 13–14 February 2023.
71. *ISO/IEC 31000:2018*; Risk Management—Guidelines. ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/65694.html (accessed on 5 February 2025).
72. *ISO/IEC 42001:2023*; Information Technology—Artificial Intelligence—Management System. ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/81230.html (accessed on 5 February 2025).
73. *ISO/IEC 23894:2023*; Information Technology—Artificial Intelligence—Guidance on Risk Management. ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/77304.html (accessed on 5 February 2025).
74. *ISO/IEC 27005:2022*; Information Security, Cybersecurity and Privacy Protection—Guidance on Managing Information Security Risks. ISO: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/80585.html (accessed on 5 February 2025).
75. Hwang, C.-L.; Yoon, K. Methods for Multiple Attribute Decision Making. In *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey*; Hwang, C.-L., Yoon, K., Eds.; Springer: Berlin/Heidelberg, Germany, 1981; pp. 58–191. [CrossRef]
76. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
77. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
78. Verma, S.; Dickerson, J.; Hines, K. Counterfactual explanations for machine learning: A review. *arXiv* **2020**, arXiv:2010.10596.
79. Kohavi, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. AAAI. Available online: https://aaai.org/papers/kdd96-033-scaling-up-the-accuracy-of-naive-bayes-classifiers-a-decision-tree-hybrid/ (accessed on 21 November 2024).
80. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv* **2018**, arXiv:1810.01943.
81. Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html (accessed on 26 February 2025).
82. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html (accessed on 26 February 2025).
83. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*; Flach, P.A., De Bie, T., Cristianini, N., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7524, pp. 35–50. [CrossRef]
84. Kamiran, F.; Karim, A.; Zhang, X. Decision Theory for Discrimination-Aware Classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 924–929. [CrossRef]
85. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 325–333. Available online: https://proceedings.mlr.press/v28/zemel13.html (accessed on 26 February 2025).
86. Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv* **2018**, arXiv:1711.05144.

87. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; ACM: Sydney, NSW, Australia, 2015; pp. 259–268. [CrossRef]
88. Sanderson, C.; Douglas, D.; Lu, Q. Implementing responsible AI: Tensions and trade-offs between ethics aspects. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023.
89. Sanderson, C.; Schleiger, E.; Douglas, D.; Kuhnert, P.; Lu, Q. Resolving Ethics Trade-Offs in Implementing Responsible AI. *arXiv* **2024**, arXiv:2401.08103.
90. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *arXiv* **2023**, arXiv:2306.08302.
91. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv* **2023**, arXiv:2308.08155.