Intelligent Decision Technologies 18 (2024) 837–854
DOI 10.3233/IDT-240366
IOS Press

Exploring the landscape of trustworthy artificial intelligence: Status and challenges

Gregoris Mentzas[a,*], Mattheos Fikardos[a], Katerina Lepenioti[a] and Dimitris Apostolou[b]
[a]*Information Management Unit, Institute of Communication and Computer Systems (ICCS), School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Athens, Greece*
[b]*Department of Informatics, University of Piraeus, Athens, Greece*

Received 21 March 2024
Accepted 28 March 2024

Abstract. Artificial Intelligence (AI) has pervaded everyday life, reshaping the landscape of business, economy, and society through the alteration of interactions and connections among stakeholders and citizens. Nevertheless, the widespread adoption of AI presents significant risks and hurdles, sparking apprehension regarding the trustworthiness of AI systems by humans. Lately, numerous governmental entities have introduced regulations and principles aimed at fostering trustworthy AI systems, while companies, research institutions, and public sector organizations have released their own sets of principles and guidelines for ensuring ethical and trustworthy AI. Additionally, they have developed methods and software toolkits to aid in evaluating and improving the attributes of trustworthiness. The present paper aims to explore this evolution by analysing and supporting the trustworthiness of AI systems. We commence with an examination of the characteristics inherent in trustworthy AI, along with the corresponding principles and standards associated with them. We then examine the methods and tools that are available to designers and developers in their quest to operationalize trusted AI systems. Finally, we outline research challenges towards end-to-end engineering of trustworthy AI by-design.

Keywords: Trustworthy artificial intelligence, responsible AI, human-AI, trust, principles, methods

1. Introduction

Artificial Intelligence (AI) has permeated every aspect of daily life and is fundamentally altering the landscape of business, economy, and society, redefining interactions and connections among stakeholders and citizens [89,43]. Organizations utilize AI advancements to enhance predictions, refine products and services, foster innovation, boost productivity and efficiency, and reduce costs, among various other advantageous applications. Many forecasts from market analysts estimate considerable increases in investments in AI software that are expected to reach approximately $300 billion in 2027 [70,54].

It is essential to emphasize, however, that the utilization of AI also presents significant risks and obstacles, prompting concerns regarding the trustworthiness of AI systems, encompassing data, algorithms, and applications [58]. Instances of biased, discriminatory, manipulative, unlawful, or human rights-violating AI deployments have exacerbated these concerns, leading to low levels of trust and acceptance. as evidenced in e.g. a recent study of more than 17,000 people from 17 countries globally [57]. The study found that while individuals express greater confidence in the ability of AI systems to deliver accurate and dependable

[*]Corresponding author: Gregoris Mentzas, School of Electrical and Computer Engineering, National Technical University of Athens, Patission 42, 10682 Athens, Greece. E-mail: gmentzas@mail.ntua.gr.

results and offer valuable services, they doubt the safety, security, and fairness of AI systems, as well as their commitment to upholding privacy rights. Moreover, although people acknowledge the numerous advantages of AI, only half of the respondents believe that these benefits outweigh the associated risks. To address these concerns, the study found that individuals expect regulatory measures to be implemented for AI.

Indeed, recently various governmental bodies have issued regulations and principles for trustworthy AI systems. As an example, within its digital strategy, the European Union enacted the AI Act [48] to govern AI, aiming to create improved conditions for the advancement and application of this technology. Similarly, the US White House issued an executive order on AI [135], setting forth fresh standards to enhance AI safety and security, safeguard privacy, and promote equity and civil rights. In addition, during the AI Safety summit in November 2023 more than 28 countries from across the globe agreed to the Bletchley Declaration on AI safety [4], which establishes a shared understanding of the opportunities and risks posed by frontier AI and the need for governments to work together to meet the most significant challenges.

Numerous companies, research institutions, and public sector organizations have released principles and guidelines aimed at promoting ethical and trustworthy AI practices. One study analysed 84 documents on ethical principles for AI [77], while another examined and compared 22 guidelines, highlighting overlaps but also omissions [61].

However, despite this proliferation of guidelines and frameworks from different organizations, it is still a challenge to implement and operationalize trustworthy AI in practice due to its complexities [25]. The implementation and operationalization of ethical AI encompass various facets in both theoretical and practical research. This includes the design, development, deployment, testing, and evaluation of approaches, all of which are underpinned by advanced AI techniques [91,38]. Recently software toolkits, approaches, and algorithms have been developed in order to support the assessment and enhancement of several trustworthiness attributes, such as fairness, explainability, etc. [80,88,137,119,24].

The present paper aims to explore this evolution in supporting the trustworthiness of AI systems. Our analysis starts from an overview of the properties of trustworthy AI, and the related principles and standards. We then examine the various methods and tools that are available to designers and developers in their quest to operationalise trusted AI systems. Finally, we outline research challenges towards end-to-end engineering of trustworthy AI by-design.

2. Trust in AI and trustworthy AI systems

2.1. Trust in AI systems

The attitudes of users toward AI systems are important in real-world AI applications. The level of trust that final end-users put in an AI system directly impacts the degree of adoption of the system. “Trust” is a multifaceted term with diverse definitions across different scientific disciplines, including psychology, sociology, economics, and computer science. Presently, there exists no standardized definition of trust [2]. Related research has found more than 300 definitions in various research areas [125]. Trust can take various forms; for example, interpersonal trust has been described as “if A [the trustor] believes that B [the trustee] will act in A’s best interest, and accepts vulnerability to B’s actions, then A trusts B” [75].

Nevertheless, when trust is directed towards a technological artifact rather than interpersonal relationships, the extent to which individuals place trust in technology hinges upon their beliefs concerning its technical attributes. This distinction has sparked debate, suggesting that technology (and consequently an AI system) cannot be trusted but only relied upon – thus it has been argued that we can only talk about the reliability and not the trustworthiness of technology [75]. Related literature argues that this objection is raised by recognizing the ‘duality of trust’ in technology, where humans rely on the technology itself and trust the technology supplier, hence trust in technology is shaped by perceptions of its functionality, utility, and reliability [40,96]. Therefore, the attribute of trustworthiness implies that the use of an AI system that is deserving of trust is based on reliable evidence.

In contrast to other technological artifacts, AI systems present a peculiarity: they employ machine learning to recognise patterns in the training data which are then used to generate algorithms for supporting or making decisions. Since these algorithms were not explicitly developed by humans but depend on the training data, it may be the case that end users assign intention to the AI systems. Arguments have been made that viewing AI systems as possessing intentions runs the risk of treating them as moral entities, implying that they bear ethical responsibility for their decisions and actions [78]. This perspective obscures the ethical obligations of AI developers and could potentially enable developers to evade responsibility and accountability for the systems they create. While AI systems may bear causal responsibility for decisions or actions, it must be clarified that it is the AI developers who are ethically accountable for them [78].

The definition of trustworthiness of AI according to the International Standards Organization is “the ability to meet stakeholders’ expectations in a verifiable way” [72]. This definition underscores the unique characteristics of AI systems, particularly their potential autonomy and their complex interactions with the social environment. The inherent uncertainty in the impact of AI systems emphasizes the need for their trustworthiness, to ensure that they act in alignment with the expectations of their users and the society at large.

Trust in AI systems is considered “layered” [83,115] since one has to consider all layers: the trust of data [81], the trust of technology [129], the trust of humans supervising or relying on it [27], the trust of organizations developing and deploying it [85], and finally the trust of the bodies regulating it [56].

Therefore, it becomes crucial to gain a clear understanding of the complexities underlying the trust dynamics between AI systems and their users, along with the prerequisites for crafting and implementing systems that demonstrate trustworthy attributes. The questions then become (i) which are the properties of trustworthy AI; and (ii) how can these properties be verified systematically?

2.2. *Properties of trustworthy AI systems*

According to the OECD, trustworthy AI refers to AI systems that adhere to the OECD AI Principles. These principles encompass AI systems that uphold human rights and privacy, are equitable, transparent, explainable, resilient, secure, and safe while ensuring accountability among all involved actors [111]. Representing the first AI standard at the intergovernmental level, these principles were endorsed in May 2019 by the 37 OECD member countries and five non-member countries, and further backed by the G20 in June 2019 [111]. The OECD AI Principles advocate for five values-based principles to guide the responsible management of trustworthy AI: inclusive growth, sustainable development, and well-being; human-centric values and equity; transparency and explainability; resilience, security, and safety; and accountability.

On the other hand, the European Union and the United States have developed a joint roadmap on evaluation and measurement tools for trustworthy AI and risk management [140]. This roadmap takes practical steps to advance trustworthy AI and uphold the shared commitment of the EU and the US to the OECD Recommendation on AI. The roadmap aims to establish a shared repository of metrics for assessing the trustworthiness of AI and methods for managing risks. Additionally, it has the potential to facilitate the development of collaborative strategies within international standards organizations focusing on Artificial Intelligence. The roadmap is informed by the efforts of the National Institute of Standards and Technology of the US Department of Commerce which has already developed the NIST AI Risk Management Framework and its related guides and tools [108] and the work related to the EU AI Act and he related deliverables of the EU High-Level Expert Group [47], such as the ALTAI Assessment List for Trustworthy AI [46].

The terminology of trustworthiness in AI is quite vague, ambiguous, and sometimes overlapping terms are used for various characteristics of trustworthiness which leads to confusing results. For example, another term that is commonly used refers to 'responsible AI', which addresses similar concepts, methods, and tools; see e.g. [10,13,18,39,124]. Actually, it has been argued that there is a risk that policymakers and the technical community could find themselves in what is referred to as the 'Inigo Montoya problem,' referring to a character in the novel and film 'The Princess Bride', specifically a scene in which Inigo Montoya understands the word 'inconceivable' differently than the main character, Vizzini, and says "You keep using that word, I do not think it means what you think it means" [118]. Similarly, the policymaking and technical communities are ascribing different meanings to the same term, leading to obvious problems in ensuring the trustworthiness of AI and its application. This may be due to several reasons ranging from the fact that related research is expanding and evolving rapidly, to the very nature of trustworthy AI, i.e. an emerging, multifaceted concept that is not bound within a singular research area.

It is quite fortunate that one of the first outcomes of the EU-US joint roadmap is an initial draft of terminology and taxonomy for Artificial Intelligence [45]. The draft terminology issued a list of 65 key AI terms essential to understanding risk-based approaches to AI. According to this terminology, the definition of trustworthy AI (based on a combination of the EU HLEG ALTAI [46] and the NIST AI RMF 1.0 [108]) should be lawful, ethical, and robust.

In parallel to the NIST AI RMF and the implementation of the EU AI Act, the International Standards Organization has recently (December 2023) published the ISO 42001 [73], which – similarly to the NIST AI RMF – is voluntary and is intended to be adaptable and scalable based on the needs of the organization that adopts it and its size. ISO 42001 outlines the requirements and offers guidance for instituting, executing, sustaining, and enhancing an AI management system within an organization's framework. It can be regarded as a complement to the NIST AI RMF, since both ISO 42001 and NIST address policies and governance procedures that organizations should contemplate to comprehensively oversee AI systems.

3. Principles, standards, methods, and tools

3.1. Principles for trustworthy AI

Currently, there are more than 1,000 AI policy initiatives from 70 countries around the world, as well as over 170 emerging initiatives that address topics of reliable and trustworthy AI [113]. These have been registered in the database of national AI Policies and strategies of the OECD.AI Policy Observatory.

Table 1 presents some well-known approaches and guidelines from international as well as national organizations, while a recent evolution at the international scene is the establishment of a "Global Challenge to Build Trust in the Age of Generative AI" [112] by organizations like the OECD, the Global Partnership on AI, the IEEE Standards Association and UNESCO.

3.2. Standards for trustworthy AI

There are also some significant efforts towards the standardization of the trustworthy elements by standards organizations and certification societies, while the European Commission issued a new standardisation request to support the recent EU AI Act [17]. Table 2 presents the main standardization approaches of organizations like ISO, IEEE, etc.

3.3. Methods for trustworthy AI

Although the policy initiatives, principles, and guidelines are important, designers and developers of

Table 1
International and national approaches principles and guidelines for trustworthy AI

Ref.	Document	Type	Description
[111]	OECD Catalogue of Tools and Metrics for Trustworthy AI	Catalogue of tools and metrics	Contains tools (technical, procedural and educational – submitted by their creators) and metrics and benchmarks for trustworthiness across the AI lifecycle.
[142]	UNESCO	Recommendation on ethical issues	Recommendation that addresses ethical issues related to AI based on a holistic, comprehensive, multicultural and evolving framework.
[59]	Global Partnership on AI, Working Group on Responsible AI	Projects on governance, algorithms, etc.	Working group on Responsible AI with a mandate to foster and contribute to the responsible development, use and governance of human-centred AI systems.
[46]	European Commission, Assessment list for trustworthy artificial intelligence	Self-assessment list	A detailed assessment list (with web tool) that supports organisations to self-assess the trustworthiness of their AI systems, according to the requirements outlined by the Ethics Guidelines for Trustworthy Artificial Intelligence (AI).
[108]	US National Institute of Standards and Technology, AI RMF	Risk management framework	A risk management framework to better manage risks to individuals, organizations, and society associated with artificial intelligence (AI). Aims to incorporate trustworthiness into the design, development, use, and evaluation of AI products, services, and systems.
[26]	China Academy of Information and Communications Technology (CAICT)	Framework	Framework of the AI governance consensus with a focus on trustworthy AI technology, industry, and industry practices. Analysis of paths to achieve reliable, transparent, and explainable AI and suggestions for developing trustworthy AI.
[141]	UK Department for Science, Innovation and Technology	Case studies	Annotated list of cases studies and a range of technical, procedural and educational approaches, illustrating how a combination of different techniques can be used to promote responsible and trustworthy AI.
[133]	TAILOR network	Network of research centres	Network of research centres aiming to develop the scientific foundations for trustworthy AI through the integration of learning, optimisation and reasoning.
[35]	Australia’s CSIRO, AI Ethics Framework	Principles and framework	Ethics framework, toolkit and governance principles and measuresI.
[23]	CERTAIN, Centre for European Research in Trusted AI	Techniques and certification approach	Collaborative initiative focused on researching, developing, deploying, standardizing, and promoting Trusted AI techniques, with the aim of providing guarantees for and certification of AI systems.
[29]	Canadian Confiance.ia	Methods and tools	Consortium that develops methods and tools to generate solutions to the challenges linked to the industrialization of sustainable, ethical, safe and responsible AI.
[130]	Singapore’s AI Verify framework	Methods and toolkit	An AI governance testing framework and a software toolkit. The testing framework aims to help organisations validate the performance of AI systems.
[3,28]	French Confiance.ai	Methods and toolkit	Industrial and academic program developing an engineering environment for trustworthy application in critical systems.

Table 2
Standards and certification programs for trustworthy AI

Ref.	Organisation/project	Type	Description
[72]	ISO/IEC TR 24028 2020	Standard	Surveys approaches to establish trust in AI systems through transparency, explainability, controllability, etc., engineering pitfalls and typical associated threats and risks to AI systems and approaches to assess and achieve trustworthiness characteristics of AI systems.
[73]	ISO/IEC 42001 2023	Standard	International standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations, ensuring responsible development and use of AI systems.
[74]	ISO/IEC DIS 42005	Standard (draft)	International standard that provides guidance for organizations performing AI system impact assessments. It includes considerations for how and when to perform such assessments and at what stages of the AI system lifecycle, as well as includes how this process can be integrated into an organization's AI risk management system.
[71]	IEEE Standards Association, IEEE CertifAIEd	Certification program	Certification program for assessing ethics of Autonomous Intelligent Systems (AIS) for their conformity to ethical privacy, transparency, accountability, and algorithmic bias criteria.
[66]	German ZERTIFIZIERTE KI project	Assessment criteria and tools	AI assessment criteria and assessment tools, safeguards to mitigate AI-related risks, development of AI standards and investigation of new business models and markets for AI assessments and certification.
[30,116]	Fraunhofer Institute for Intelligent Analysis and Information Systems	Audit areas for AI certification	Formulates AI-specific audit areas for trustworthy use of artificial intelligence, providing the basis for an AI audit catalogue.
[37,36]	UK Department for Science, Innovation and Technology	Assurance techniques and ecosystem	Roadmap at the role of AI assurance in ensuring trusted and trustworthy AI. Description on how assurance engagements can build justified trust in AI systems. Structure of assurance engagements and assurance tools relevant to AI and their applications for ensuring trusted and trustworthy AI systems.

AI systems lack any clear actionable instructions on how to practically implement them [102]. The vague terminology and abstract nature of some of the principles and policy guidelines have not been helpful to data scientists, machine learning engineers, and designers who are the ones involved in operationalizing these principles in practice [21]. Hence several structured procedural frameworks, methods, and auditing procedures have been developed to assist organizations in their efforts to design, develop, and audit the trustworthiness of their AI systems; Table 3 presents indicative approaches.

The UC Berkeley Center for Long-Term Cybersecurity published a taxonomy of 150 trustworthiness properties and mapped them across the AI lifecycle stages of the NIST RMF, thereby creating a resource and tool for organizations developing AI, as well as for standards-setting bodies, policymakers, independent auditors, and civil society organizations working to evaluate and promote trustworthy AI [109].

This large number of diverse properties of trustworthiness and the different approaches that allow e.g. mitigation of negative effects, increase the need for structured approaches to the design and development of trustworthy AI systems. A recent initiative [107] that aimed at aiding the selection of the most suitable framework examined and categorized over 40 existing responsible AI frameworks. These frameworks were then mapped onto a matrix, facilitating organizations in comprehending, selecting, and implementing responsible AI in alignment with their specific requirements. The matrix comprises two dimensions: the user dimension, representing individuals responsible for implementing frameworks within organizations involved in building or utilizing AI systems, and the utility dimension, which organizes frameworks into three categories: components, AI lifecycle, and trustworthiness characteristics.

Another review outlined in [117] analyzes over 100 frameworks, process models, and proposed solutions and tools aimed at facilitating the transition from principles to implementation. Building upon the work in [105], this analysis underscores the emphasis of existing approaches on a select few ethical concerns such as explainability, fairness, privacy, and accountability. It also introduces a more refined segmentation of the AI development process and identifies areas necessitating further scrutiny from researchers and developers.

3.4. Software toolkits for trustworthy AI

The abundance of conceptual principles, guidelines, and methods has been recently accompanied by many concrete software tools that attempt to address the need to move from ‘what’ to ‘how’, i.e. to move beyond ethical AI guidelines to concrete operational mandates and tools that enable better oversight mechanisms in the way AI systems are developed and deployed.

Various survey papers review the related technologies and tools. For example, [105] reviews tools and methods in order to help translate principles into practice, while [86] introduces a framework that consolidates the existing fragmented approaches to trustworthy AI into a unified, systematic approach. This approach encompasses the entire lifecycle of AI systems, spanning from data acquisition to model development, system development and deployment, and ultimately to continuous monitoring and governance. [138] focuses on four categories of system properties that are considered instrumental in achieving the policy objectives of AI trustworthiness, namely fairness, explainability, auditability and safety & security (FEAS). They further review the main technologies and tools with respect to these four properties, for data-centric as well as model-centric stages of the machine learning system life cycle. The authors of [88] concentrate on six dimensions crucial for attaining trustworthy AI: (i) Safety & Robustness, (ii) Non-discrimination & Fairness, (iii) Explainability, (iv) Privacy, (v) Accountability & Auditability, and (vi) Environmental Well-being. For each dimension, they assess the associated technologies, outline their real-world applications, and explore the corresponding and conflicting interactions among these various dimensions. On the other hand, [80] analyses trustworthiness requirements (fairness, explainability, accountability, reliability, and acceptance) adopting a human-centered approach by examining different levels of human involvement in making AI systems trustworthy.

Table 3
Procedural methods for trustworthy AI

Ref.	Document	Type	Description
[114,51,103]	Oxford Internet Institute, capAI	Procedural method	A procedure for conformity assessment of AI systems aligned with the EU's proposed AI Act. Intended to support independent assessment and provide guidance on translating high-level ethics principles into verifiable criteria for trustworthy AI.
[146,147]	Z-Inspection	Evaluation process	Process used to evaluate the trustworthiness of AI systems at different stages of the AI lifecycle. It focuses on the identification of ethical issues and tensions through the elaboration of socio-technical scenarios and uses the EU HLEG guidelines for trustworthy AI.
[44]	Etami consortium	Guidelines and auditing procedures	Actionable guidelines to develop legal, trustworthy, and ethical AI. The focus is put on quality-centric lifecycle models for AI systems, legal compliance, and AI auditing practices.
[50]	AI Ethics Impact Group	Guidelines and procedure	Guidance on how to incorporate values into algorithmic decision-making using the VCIO (Values, Criteria, Indicators, Observables) model combined with a context-dependent risk assessment.
[14,15]	TAII Framework	Guidelines and procedure	Business management framework to initiate trustworthy AI implementations by the analysis of ethical inconsistencies and dependencies for a planned AI system. The TAII Framework considers these dependencies: corporate values, business models, and common good.

Table 4
Software toolkits for trustworthy AI

Ref.	Organisation	Software	Description
[6]	AI Verify Foundation	https://github.com/IMDA-BTG/aiverify	Single integrated toolkit that can perform technical tests on common supervised learning classification and regression models for most tabular and image datasets.
[68]	IBM research trustworthy AI	https://github.com/Trusted-AI	Various projects available on Linux Foundation AI Trusted AI organisation e.g AI Fairness 360, AI Explainability 360, etc.
[69]	IBM product watsonx.governance	Not openly available	watsonx.governance^{TM}employs software automation to strengthen an organisation's ability to mitigate risks, manage regulatory requirements and address ethical concerns.
[134]	Google TensoFlow Responsible AI	https://github.com/tensorflow	Responsible AI practices (e.g. fairness, privacy, interpretabiloity) integrated in the ML workflow using TensorFlow.
[100,101]	Microsoft Responsible AI	https://github.com/microsoft/responsible-ai-toolbox tutorials & walkthroughs https://github.com/microsoft/responsible-ai-workshop	Suite of tools providing a collection of model and data exploration and assessment user interfaces and libraries that empower developers and stakeholders of AI systems to develop and monitor AI more responsibly. Microsoft also provides a series of hands-on tutorials for developers and data scientists.
[123,122]	SAS product Viya Platform	Not openly available	SAS Viya AI and data analytics platform with AI-based automation produces outcomes that are repeatable, reliable, explainable and compliant.
[34]	Data Robot	Not openly available	DataRobot's enterprise AI platform incorporates features and tools that make trustworthy AI accessible and standardized.
[32,33]	DataIku Govern	https://github.com/dataiku	Dataiku governance framework features a centralized monitoring capability and integrated MLOps to close the governance loop after models are deployed into production.
[22]	Captum	https://github.com/pytorch/captum	Captum provides algorithms that allow developers to understand which features are contributing to a model's output.
[8]	Alexandra Institute	https://github.com/alexandrainst/responsible-ai	Knowledge base for responsible AI.
[65,31]	Holistic AI library	https://github.com/holistic-ai/holisticai Documentation https://holisticai.readthedocs.io/en/latest/	The Holistic AI library is an open-source tool to assess and improve the trustworthiness of AI systems. Currently, it offers a set of techniques to easily measure and mitigate bias and in the future it will be extended to include tools for efficacy, robustness, privacy and explainability.
[139]	Trustible solution	Not openly available	Responsible AI Governance platform, a turnkey solution to maximize trust and facilitate AI governance.

In addition to academically available research efforts, major technology companies have started providing technologies and toolkits to support trustworthy AI. Table 4 outlines the efforts of major well-known corporations, some of which (like IBM and Microsoft) provide open-source versions of their toolkits.

The first is the AI Verify Foundation [6], a not-for-profit subsidiary of IMDA, the Infocommunications Media Development Authority of Singapore. This initiative seeks to leverage the collective expertise and efforts of the global open-source community to create AI testing tools that promote responsible AI practices. Key members of this foundation include industry giants such as Google, IBM, Microsoft, RedHat, Aicadium, Salesforce, among others. The foundation is responsible for the development of AI Verify, a framework and software toolkit designed for AI governance testing. AI Verify validates the performance of AI systems based on a set of principles and aligns with AI governance frameworks such as those established by the European Union, OECD, and Singapore.

The second is the LF AI and Data Foundation, a Linux Foundation project that supports and sustains open-source projects within AI and the data space [87]. Of relevance is the Trusted-AI that hosts LF AI Foundation projects in the category of Trusted and Responsible AI. Among them are IBM's toolkits such as AI Fairness 360, AI Explainability 360, Adversarial Robustness 360, AI Privacy 360, etc.

While all these tools primarily concentrate on evaluating the trustworthiness of the AI system itself, there are recent research endeavours that redirect attention towards assessing the perceived trustworthiness of the development process. The rationale behind this shift is that while trustworthy AI is defined by system requirements, its practical implementation necessitates an understanding of its connection to specific measures throughout the development process. For example, [64] presents a concept for establishing a trustworthy development process for AI systems, introducing a framework derived from a semi-systematic analysis of AI governance activities. This framework aims to identify obligations and measures necessary to meet established AI ethics requirements and align them with the AI development lifecycle. Another effort in [121] focuses on requirements engineering and examines the applicability of ethical AI development frameworks for performing effective requirements engineering during the development of trustworthy AI systems.

4. Challenges and directions for trustworthy AI systems

Although much work has been done, there are still considerable research challenges to be tackled to safely guarantee trustworthy AI systems. In the following, we examine five such challenges: (i) the need to shift from human-in-the-loop approaches to modelling and supporting teaming of humans and AI systems; (ii) the quantification and monitoring of key trustworthiness indicators; (iii) the potential that recent trends in neuro-symbolic AI may generate for human-centric trustworthy AI; (iv) the methods and tools for supporting end-to-end trustworthy AI system engineering; and (v) the need to explicitly address the complexities and particularities of generative AI.

4.1. *Shift from human-in-the-loop to human-AI teams*

It has been argued that in order to guarantee trustworthy, responsible, and ethical AI, researchers and practitioners have to adopt a Human-Centered AI approach (HCAI) and consider hybrid human-AI intelligence [7]. This approach strives to develop AI solutions that mitigate discrimination and uphold fairness and justice, while also accurately reflecting human intelligence. It places explicit emphasis on human factors design to ensure that AI solutions are explainable, understandable, useful, and user-friendly [127,128,145]. For example, [131] introduces the Human-Centered Trustworthy Framework which aims to elucidate the connections among user trust, socio-ethical considerations, technical and design elements, and user attributes. This framework is designed to offer AI providers, designers, and other stakeholders, straightforward guidelines for integrating user trust considerations into AI design.

A research challenge lies in transitioning from the ‘human-in-the-loop’ approach to the ‘human-AI teaming’ approach. In this paradigm, AI systems collaborate with humans to accomplish tasks, often within larger teams comprising both humans and AI systems. These AI systems may exhibit varying levels of autonomy, operate in different contexts, and handle diverse tasks, leading to a broad design spectrum to consider. Consequently, designing and deploying AI systems that effectively collaborate with humans pose considerable challenges, including ensuring adequate levels of AI transparency and explainability to facilitate human situational awareness, and supporting seamless collaboration and coordination between humans and AI systems [41,49].

Recent research has been exploring the role of agent reliability on human trust, the methods of communicating intent between human and AI agents, and ways that AI agents either work together with humans or work as trainers of humans [42]. New approaches that pave the way for innovative research in this direction include for example considering coordination in hybrid teams of humans and autonomous agents in many-to-many situations (multiple humans and multiple agents) with the use of trustworthy interaction patterns [97], modelling the maturity of collaborative human-AI ecosystems [106] separating the design choices in terms of the different decision tasks and evaluating the efficacy, usability, and reliance of approaches [84].

Such approaches hold significant potential, particularly in critical domains such as healthcare, transportation (including driving and aviation), military operations, and search and rescue missions. In these fields, the development of effective methods for seamlessly integrating AI with human operations is of utmost importance [104].

The path towards establishing trustworthy human-AI teaming, spanning from initial conception to the formation of high-performing teams, encompasses a wide array of areas including bi-directional situational awareness, human-AI interaction, intelligent decision-making, and human-AI operations [20]. Furthermore, it is noted that this research landscape is defined more by open research questions than by existing knowledge. Many of the questions that researchers will confront along this path are likely yet to be posed, let alone answered [20].

4.2. *Quantify and monitor trustworthy indicators*

A critical topic that generates interesting research relates to the quantification and monitoring of indicators of AI trustworthiness [144]. Although there may exist Key Performance Indicators (KPIs) and metrics for each separate attribute of trustworthiness like fairness or privacy, the challenge is twofold: (a) to adopt a holistic approach for measuring most (if not all) of trustworthiness characteristics in an integrated manner; and (b) to explicitly take into the peculiarities and specific features of the AI system in-use.

The first challenge centers on quantifying the trustworthiness of AI systems, builds upon existing metrics and indices and assigns scores to the various attributes. For example, [67] focuses on supervised machine and deep learning models, develops an algorithm that considers twenty-three metrics grouped into four pillars of trusted AI (fairness, explainability, robustness, and accountability), and aggregates the metrics to calculate a global trustworthiness score. Likewise, research conducted within the French Confiance.ai program [93] delineates various attributes contributing to the concept of trustworthiness. It delves into each attribute to identify associated Key Performance Indicators (KPIs), assessment methods, or control points, and establishes an aggregation methodology for these attributes.

The second challenge refers to the need to consider the context of the use of the AI system. Contextual factors influencing trustworthiness attributes include the criticality level of the application, the domain of the application, the anticipated use of the AI system, and the involved stakeholders, among others. This implies that in different contexts, certain attributes may take precedence, while additional attributes may be introduced to the list. For instance, a medical imaging system tailored for medical professionals may entail distinct trustworthy requirements compared to a human resource management application. In this case, multi-criteria decision support methods are well suited for assessing the various characteristics and may provide the appropriate instruments for aggregating individual preferences and scores; see e.g. [9,90,94].

4.3. Exploit neuro-symbolic AI

Many researchers have identified the need to integrate well-founded knowledge representation and reasoning with deep learning [62,126]. This has spurred the development of neuro-symbolic computing, which has emerged as a promising area of research aiming to integrate robust learning within neural networks with symbolic knowledge representation and logical reasoning [63]. This trend seeks to leverage the parallels often drawn by AI researchers between Kahneman's investigations into human reasoning and decision-making, as detailed in his book 'Thinking, Fast and Slow' [79], and the concept of 'AI systems 1 and 2'. In this model, deep learning would correspond to AI system 1, responsible for intuitive and rapid decision-making, while symbolic reasoning would correspond to AI system 2, handling deliberate and logical reasoning [53]. Although some basic ingredients of neuro-symbolic AI have already been suggested, there are still many outstanding challenges, like the choice of an appropriate language, the need for standard benchmarks, etc. [53].

The adoption of neuro-symbolic AI holds promise as a potential solution for enhancing the reliability, robustness, and trustworthiness of AI systems. Furthermore, neuro-symbolic AI can aid in enhancing integration by addressing bias, improving data quality, aligning AI with human values, and furnishing human-comprehensible explanations for AI-generated predictions. A recent review [99] yielded a total of 54 papers that employed neuro-symbolic methods with an emphasis on trustworthiness (and identified a clear focus on interpretability), while [120] claims that neuro-symbolic AI can assist operations across critical domains with high assurance and trust by helping to provide robustness to adversarial perturbations and assurance by analysing heterogeneous evidence towards safety and risk assessments. Furthermore, [55] seeks to showcase the suitability of neuro-symbolic AI for crafting trustworthy AI by introducing the CREST framework. This framework illustrates how Consistency, Reliability, user-level Explainability, and Safety are established through neuro-symbolic methods, which leverage both data and knowledge to meet the demands of critical applications such as healthcare and well-being.

Although the review in [99] highlights a noteworthy dedication of trustworthy neuro-symbolic AI to enhancing interpretability, it also identified a noticeable portion of work aimed at improving other aspects of trustworthiness, which underscores an opportunity for future research to extend the benefits of neuro-symbolic AI to other critical dimensions of trustworthiness.

4.4. Develop and adopt trustworthy AI system engineering methods

To guarantee the trustworthiness of their AI system, designers and developers should not consider it as a requirement to be satisfied ex post, after their system is deployed; on the contrary, the main attributes of trustworthiness need to be addressed from the early start of the conception and design of the system. The engineering of trustworthy AI systems needs to consider the trustworthiness requirements of their stakeholders. This is not the current status quo. For instance, machine-learning applications are typically defined according to optimization and efficiency criteria rather than considering quality requirements that align with the needs of stakeholders.

Rigorous specification techniques for the development and deployment of AI applications are essential. These techniques are being explored within the U.K. Research and Innovation (UKRI) Trustworthy Autonomous Systems (TAS) program. This program undertakes cross-disciplinary fundamental research to guarantee that systems are safe, reliable, resilient, ethical and trusted [1]. This work gives rise to notable research challenges, including the formalization of human-understandable knowledge to render it interpretable by machines; the specification and modelling of human behaviour, intent, and mental state; and the modeling of social and ethical norms pertinent to human-AI interaction [1].

Another notable approach in this direction is followed by the French Confiance.ai project [28] which adopts model-based systems engineering for the assessment of trustworthiness from the early stages of design up to the deployment and operation of the AI system [12,11,93]. The approach is an adaptation of the Arcadia method [143] and hence is built around four perspectives: Operational Analysis (the engineering methods and processes, the operational need around the Trustworthiness Environment), System Analysis (the functions of the Trustworthiness Environment), Logical Architecture and Physical Architecture (abstract and concrete resources of the Trustworthiness Environment). The system of interest in this case is the Trustworthy Environment, the tooled workbench to be delivered by the Confiance.ai research program and the overall ambition is to obtain an applicable end-to-end engineering method [12].

4.5. Towards trustworthy and responsible generative AI

The recent explosion of foundation models and generative AI models and applications and their deployment across a wide spectrum of industries is considered to have a tremendous impact on productivity and the way of work. For example, McKinsey and Company estimate that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion annually to the global economy – by comparison, the United Kingdom's entire GDP in 2021 was $3.1 trillion [95]. Nevertheless, generative AI presents various ethical and social concerns. Problems such as the absence of interpretability, bias and discrimination, privacy breaches, lack of model robustness, dissemination of fake and misleading content, copyright infringement, plagiarism, and environmental consequences have been linked to the training and inference processes of generative AI models.

Installing guard rails with appropriate tooling has become imperative, in order to address these problems and has expanded the requirements for AI trustworthiness. New requirements include, for example, verifying and validating GenAI models before they become available, or reporting in a more interoperable and standardized way on the testing that has been done to improve transparency and accountability.

In response to these concerns, the US National Institute of Standards launched a Generative AI working group to devise a profile of AI Risk Management Framework (RMF) tailored specifically for generative AI. This initiative aims to establish four sets of guidelines covering pre-deployment verification and validation of generative AI models, digital content provenance, incident disclosure, and governance of generative AI systems [110]. The AI Verify Foundation proposed a novel model AI governance framework for generative AI, aiming to introduce a systematic and balanced approach to address the concerns associated with generative AI [5]. This framework outlines nine dimensions to be collectively considered to cultivate a trusted ecosystem: accountability, data, trusted development and deployment, incident reporting, testing and assurance, security, content provenance, safety and alignment R&D, and AI for social good.

Ongoing research is dedicated to aiding data scientists and machine learning developers in constructing generative AI systems that prioritize security, privacy preservation, transparency, explainability, fairness, and accountability. The objective is to mitigate unintended consequences and address compliance challenges that could pose harm to individuals, businesses, and society at large [82].

5. Conclusions

The remarkable advancements in Artificial Intelligence and its widespread integration into nearly every aspect of daily life underscore the importance of prioritizing 'trust' as a fundamental principle in the design, development, and monitoring of AI, rather than considering it optional [19]. Various national and international bodies have introduced a range of principles and regulations that AI systems must adhere to in order to earn trust. While different terms like AI validation, assessment, auditing, or monitoring may be used, the essential objective remains to ensure that AI systems function effectively within their intended operational parameters and comply with regulatory guidelines [136,132].

This paper explored the alternative approaches, standards, methods, and tools that recently emerged in efforts to enable trustworthy AI. Although much work has been done already, there are still open research issues for end-to-end AI engineering that enables ‘trust-by-design’ [92,98]. The field of AI trustworthiness assessment is rapidly evolving, new attributes and evaluation metrics may emerge and new avenues for research will be explored. Ensuring AI trustworthiness will be a critical element in our quest to develop AI for social good [52].

Acknowledgments

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’ view and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] Abeywickrama DB, Bennaceur A, Chance G, Demiris Y, Kordoni A, Levine M, et al. On specifying for trustworthiness. Communications of the ACM. 2023; 67(1): 98–109.
[2] Adams BD, Bruyn LE, Houde S, Angelopoulos P, Iwasa-Madge K, McCann C. Trust in automated systems. Ministry of National Defence. 2003.
[3] Adedjouma M, Adam JL, Aknin P, Alix C, Baril X, Bernard G, et al. Towards the engineering of trustworthy AI applications for critical systems – The Confianceai program. 2022.
[4] AI Safety Summit. The Bletchley Declaration by Countries Attending the AI Safety Summit. 1–2 November 2023.
[5] AI Verify. Proposed Model AI Governance Framework for Generative AI: Fostering a Trusted Ecosystem. 2024.
[6] AI Verify Foundation. 2024. Available from https://aiverifyfoundation.sg.
[7] Akata Z, Balliet D, De Rijke M, Dignum F, Dignum V, Eiben G, et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. 2020; 53(8): 18–28.
[8] Alexandra Institute. Available from https://alexandra.dk/about-the-alexandra-institute/.
[9] Alsalem MA, Alamoodi AH, Albahri OS, Albahri AS, Martínez L, et al. Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. Expert Systems with Applications. 2024; 246.
[10] Anagnostou M, Karvounidou O, Katritzidaki C, Kechagia C, Melidou K, et al. Characteristics and challenges in the industries towards responsible AI: A systematic literature review. Ethics and Information Technology. 2022; 24(3): 37.
[11] Awadid A, Amokrane-Ferka K, Sohier H, Mattioli J, Adjed F, Gonzalez M, Khalfaoui S. AI Systems Trustworthiness Assessment: State of the Art. In Workshop on Model-based System Engineering and Artificial Intelligence-MBSE-AI Integration. 2024.
[12] Awadid A, Robert B, Langlois B. MBSE to Support Engineering of Trustworthy AI-Based Critical Systems. In 12th International Conference on Model-Based Software and Systems Engineering. 2024.
[13] Baeza-Yates R. Introduction to Responsible AI. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 2024. pp. 1114–1117.
[14] Baker-Brunnbauer J. Trustworthy Artificial Intelligence Implementation: Introduction to the TAII Framework. Springer Nature. 2022.
[15] Baker-Brunnbauer J. TAII Framework for Trustworthy AI Systems. ROBONOMICS: The Journal of the Automated Economy. 2021; 2: 17.
[16] Barz T, Loevenich D, Poretschkin M. TAISEC & TAISEM An approach to horizontal Criteria for AI Systems Evaluation & Certification, presentation at the Trustworthy AI Standardization Workshop, organized by ZERTIFIZIERTE KI, 27th October 2023, Singapore. 2023.
[17] Becker F, Lehmann T. EU Al Act Standardization Request – The European Understanding of AI Trustworthiness,

presentation at the Trustworthy AI Standardization Workshop, organized by ZERTIFIZIERTE KI, 27th October 2023, Singapore. 2023.
[18] Benjamins R, Barbado A, Sierra D. Responsible AI by design in practice. arXiv preprint arXiv:1909.12838. 2019.
[19] Braunschweig B, Buijsman S, Chamroukhi F, Heintz F, Khomh F, Mattioli J, Poretschkin M. AITA: AI trustworthiness assessment: AAAI spring symposium. AI and Ethics. 2024.
[20] Caldwell S, Sweetser P, O'Donnell N, Knight MJ, Aitchison M, Gedeon T, et al. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. ACM Transactions on Interactive Intelligent Systems (TiiS). 2022; 12(3): 1–36.
[21] Canca C. Operationalizing AI ethics principles. Communications of the ACM. 2020; 63(12): 18–21.
[22] Captum.ai. Available from https://captum.ai.
[23] Certain-Trust. Available from https://www.certain-trust.de.
[24] Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (xai). IEEE Access. 2023; 78994–79015.
[25] Chen F, Zhou J, Holzinger A, Fleischmann KR, Stumpf S. Artificial intelligence ethics and trust: From principles to practice. IEEE Intelligent Systems. 2023; 38(6): 5–8.
[26] China Academy of Information and Communications Technology. White Paper on Trustworthy Artificial Intelligence. 2021. Available from http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf.
[27] Chiou E, Lee JD. Trusting Automation: Designing for Responsivity and Resilience. Human Factors. 2021; 137–165.
[28] Confiance.ai – the French technological research programme on trustworthy AI. Available from https://www.confiance.ai/en/.
[29] ConfianeIa. Available from https://www.confianceia.ca.
[30] Cremers AB, Englander A, Gabriel M, Hecker D, Mock M, Poretschkin M, et al. Trustworthy Use of Artificial Intelligence – priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of artificial intelligence. Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). 2019.
[31] da Costa K. Using Python to Mitigate Bias and Discrimination in Machine Learning Models, Holistic AI Blob. 2023. Available from https://www.holisticai.com/blog/using-python-to-mitigate-bias-and-discrimination.
[32] DataIku. A Lifecycle Approach for Responsible. Available from https://blog.dataiku.com/a-lifecycle-approach-for-responsible-ai.
[33] DataIku Knowledge Base – DataIku Govern. Available from https://knowledge.dataiku.com/latest/mlops-o16n/govern/index.html.
[34] DataRobot Trusted. AI 101: A Guide to Building Trustworthy and Ethical AI Systems. Available from https://www.datarobot.com/trusted-ai-101/.
[35] Dawson D, Schleiger E, Horton J, McLaughlin J, Robinson C, Quezada G, Scowcroft J, Hajkowicz S. Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia. 2019.
[36] Department for Science, Innovation & Technology. Introduction to AI assurance. Available from https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf.
[37] Department for Science, Innovation and Technology. The roadmap to an effective AI assurance ecosystem – extended version, Available from https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version.
[38] Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, de Prado ML, 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. Information Fusion. 2023.
[39] Dignum V. Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 1). Cham: Springer. 2019.
[40] Duenser A, Douglas DM. Whom to Trust, How and Why: Untangling AI Ethics Principles, Trustworthiness and Trust. IEEE Intelligent Systems, November-December. 2023.
[41] Endsley MR. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior. 2023; 140: 107574.
[42] Endsley MR, Cooke N, McNeese N, Bisantz A, Militello L, Roth E. Special issue on human-AI teaming and special issue on AI in healthcare. Journal of Cognitive Engineering and Decision Making. 2022; 16(4): 179–181.
[43] Enholm IM, Papagiannidis E, Mikalef P, Krogstie J. Artificial intelligence and business value: A literature review. Information Systems Frontiers. 2022; 24(5): 1709–1734.
[44] etami – Ethical and Trustworthy Artificial and Machine Intelligence. The open guidebook on legal, trustworthy, and ethical Artificial Intelligence. Available from https://guidebook.etami.org.
[45] EU-U.S. Terminology and Taxonomy for Artificial Intelligence. Available from https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence.
[46] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Available from https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[47] European Commission, High-level expert group on artificial intelligence, HLEG AI. Available from https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.
[48] European Parliament. EU AI Act: first regulation on artificial intelligence, Published: 08-06. 2023.

[49] Ezer N, Bruni S, Cai Y, Hepenstal SJ, Miller CA, Schmorrow DD. Trust engineering for human-AI teams. In Proceedings of the human factors and ergonomics society annual meeting. Sage CA: Los Angeles, CA: SAGE Publications. Vol. 63, No. 1, 2019. pp. 322–326.

[50] Fetic L, Fleischer T, Grünke P, Hagendorf T, Hallensleben S, Hauer M, et al. From Principles to Practice. An interdisciplinary framework to operationalise AI Ethics. AIEI AI Ethics Impact Group. Available from https://www.ai-ethics-impact.org/en.

[51] Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J, Wen Y. CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. Available at SSRN 4064091. 2022.

[52] Foffano F, Scantamburlo T, Cortés A. Investing in AI for social good: An analysis of European national strategies. AI & Society. 2023; 38(2): 479–500.

[53] Garcez ADA, Lamb LC. Neurosymbolic AI: The 3 rd wave. Artificial Intelligence Review. 2023; 56(11): 12387–12406.

[54] Garner. Invest Implications: Forecast Analysis: Artificial Intelligence Software, 2023–2027. 2023. Available from https://www.gartner.com/en/documents/4925331.

[55] Gaur M, Sheth A. Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety. AI Magazine. 2024.

[56] Gillespie N, Lockey S, Curtis C, Pool J, Akbari A. Trust in Artificial Intelligence: A global study, University of Queensland and KPMG Australia. 2023.

[57] Gillespie N, Lockey S, Curtis C, Pool J, Akbari A. Trust in Artificial Intelligence: A global study, University of Queensland and KPMG Australia. 2023.

[58] Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals. 2020; 14(2): 627–660.

[59] Global Partnership on AI. Available from https://gpai.ai/projects/responsible-ai/.

[60] Hagendorff T. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines. 2020; 30(1): 99–120.

[61] Hagendorff T. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines. 2020; 30(1): 99–120.

[62] Hitzler P, Sarker MK. (Eds.) Neuro-symbolic artificial intelligence: The state of the art. IOS Press. 2022.

[63] Hitzler P, Eberhart A, Ebrahimi M, Sarker MK, Zhou L. Neuro-symbolic approaches in artificial intelligence. National Science Review. 2022.

[64] 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(4): 904–925.

[65] Holistic AI, AI governance platform. Available from https://www.holisticai.com/ai-governance-platform.

[66] Zertifizierte KI. Available from https://www.zertifizierte-ki.de.

[67] Huertas Celdran A, Kreischer J, Demirci M, Leupp J, Sánchez Sánchez PM, Figueredo Franco M, et al. A Framework Quantifying Trustworthiness of Supervised Machine and Deep Learning Models. In SafeAI2023: The AAAI's Workshop on Artificial Intelligence Safety. 2023.

[68] IBM research. Trustworthy AI. Available from https://research.ibm.com/topics/trustworthy-ai.

[69] IBM watsonx.governance. Available from https://www.ibm.com/products/watsonx-governance.

[70] IDC. IDC Forecasts Revenue for Artificial Intelligence Software Will Reach $307 Billion Worldwide in 2027. Available from https://www.idc.com/getdoc.jsp?containerId=prUS51345023.

[71] IEEE Standards Association. IEEE CertifAIEd$^{\mathrm{TM}}$The Mark of AI Ethics. Available from https://engagestandards.ieee.org/ieeecertifaied.html.

[72] International Organization for Standardization. ISO/IEC TR 24028:2020 Information technology, Artificial intelligence, Overview of trustworthiness in artificial intelligence. Available from https://www.iso.org/standard/77608.html.

[73] International Organization for Standardization. ISO/IEC 42001:2023 Information technology, Artificial intelligence Management system. Available from https://www.iso.org/standard/81230.html.

[74] International Organization for Standardization. ISO/IEC DIS 42005 Information technology Artificial intelligence AI system impact assessment. Available from https://www.iso.org/standard/44545.html.

[75] Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

[76] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019; 1(9): 389–399.

[77] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019; 1(9): 389–399.

[78] Johnson DG, Verdicchio M. AI, agency and responsibility: The VW fraud case and beyond. Ai & Society. 2019; 34: 639–647.

[79] Kahneman D. Thinking, fast and slow. Macmillan. 2011.

[80] Kaur D, Uslu S, Rittichier KJ, Durresi, A. Trustworthy artificial intelligence: A review. ACM Computing Surveys (CSUR). 2022; 55(2): 1–38.

[81] Kennedy H. ACM TechBrief: The Data Trust Deficit, Association for Computing Machinery. 2023.

[82] Kenthapadi K, Lakkaraju H, Rajani N. Generative ai meets responsible ai: Practical challenges and opportunities. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023. pp. 5805–5806.

[83] Knowles B, Richards JT. ACM TechBrief: Trusted AI. 2024.

[84] Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023. pp. 1369–1385.

[85] Lankton N, Harrison McKnight D. What does it mean to trust facebook? Examining technology and interpersonal trust beliefs. ACM SIGMIS Database: The Data Base for Advances in Information Systems. 2011; 42(2): 32–54.

[86] Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: From principles to practices. ACM Computing Surveys. 2023; 55(9): 1–46.

[87] Linux Foundation LF AI & Data Foundation. Available from https://lfaidata.foundation.

[88] Liu H, Wang Y, Fan W, Liu X, Li Y, Jain S, et al. Trustworthy ai: A computational perspective. ACM Transactions on Intelligent Systems and Technology. 2022; 14(1): 1–59.

[89] Makridakis S. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures. 2017; 90: 46–60.

[90] Mannion P, Heintz F, Karimpanal TG, Vamplew P. Multi-objective decision making for trustworthy ai. In Proceedings of the Multi-Objective Decision Making (MODeM) Workshop. 2021.

[91] Marzouk M, Zitoun C, Belghith O, Skhiri S. The Building Blocks of a Responsible AI Practice: An Outlook on the Current Landscape. IEEE Intelligent Systems. 2023.

[92] Mattioli J, Le Roux X, Braunschweig B, Cantat L, Tschirhart F, Robert, et al. AI engineering to deploy reliable AI in industry. In 2023 Fifth International Conference on Transdisciplinary AI (TransAI). 2023. pp. 228–231.

[93] Mattioli J, Sohier H, Delaborde A, Amokrane-Ferka K, Awadid A, Chihani Z, et al. An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. AI and Ethics. 2024.

[94] Mattioli J, Sohier H, Delaborde A, Pedroza G, Amokrane-Ferka K, Awadid, et al. Towards a holistic approach for AI trustworthiness assessment based upon aids for multi-criteria aggregation. In SafeAI 2023-The AAAI's Workshop on Artificial Intelligence Safety. 2023.

[95] McKinsey & Company. The economic potential of generative AI: The next productivity frontier. 2023.

[96] Mcknight DH, Carter M, Thatcher JB, Clay PF. Trust in a specific technology: An investigation of its components and measures. ACM Transactions on Management Information Systems (TMIS). 2011; 2(2): 1–25.

[97] Meyer-Vitali A, Mulder W. Trustworthy Hybrid Team Decision-Support. In Proceedings of the First International Conference on Hybrid Human-Machine Intelligence. Workshop on the Representation, Sharing and Evaluation of Multimodal Agent Interaction, befindet sich International Conference on Hybrid Human-Artificial Intelligence (HHAI). 2022. pp. 13–17.

[98] Meyer-Vitali A. AI Engineering for Trust by Design, United Innovations. 202. Release 1. 2024. pp. 20–22.

[99] Michel-Delétie C, Sarker MK. Neuro-Symbolic methods for Trustworthy AI: a systematic review. Neurosymbolic Artificial Intelligence. 2024.

[100] Microsoft AI Lab – Responsible AI Dashboard. Available from https://www.microsoft.com/en-us/ai/ai-lab-responsible-ai-dashboard.

[101] Microsoft Resonsible AI Toolbox. Available from https://responsibleaitoolbox.ai.

[102] Mittelstadt B. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 2019; 1(11): 501–507.

[103] Mökander J, Floridi L. Operationalising AI governance through ethics-based auditing: An industry case study. AI and Ethics. 2023; 3(2): 451–468.

[104] Monaro M, Barakova E, Navarin N. Editorial special issue interaction with artificial intelligence systems: New human-centered perspectives and challenges. IEEE Transactions on Human-Machine Systems. 2022; 52(3): 326–331.

[105] Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics. 2020; 26(4): 2141–2168.

[106] Mulder W, Meyer-Vitali A. A Maturity Model for Collaborative Agents in Human-AI Ecosystems. In Working Conference on Virtual Enterprises. Cham: Springer Nature Switzerland. 2023.

[107] Narayanan M, Schoeber C. A Matrix for Selecting Responsible AI Frameworks (Center for Security and Emerging Technology, June 2023). Available from doi: 10.51593/20220029.

[108] National Institute of Stadards and Technology. NIST AI Risk Management Framework. Available from https://www.nist.gov/itl/ai-risk-management-framework.

[109] Newman J. A taxonomy of trustworthiness for artificial intelligence. UC Berkeley Center for Long-Term Cybersecurity, CLTC: North Charleston, SC, USA. 2023.

[110] NIST. Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence, Testimony of Elham Tabassi, Associate Director for Emerging Technologies to the United States House of Representatives. Available from https://www.nist.gov/speech-testimony/balancing-knowledge-and-governance-foundations-effective-risk-management-artificial.

[111] OECD. "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems" OECD Digital Economy Papers. No. 312. OECD Publishing. 2021.

[112] OECD. Call for Partners: the Global Challenge to Build Trust in the Age of Generative AI. Available from https://oecd.ai/en/wonk/global-challenge-partners.

[113] OECD. AI, powered by EC OECD. 2021. Database of national AI policies. Available from https://oecd.ai.

[114] Oxford Internet Institute, capAI capAI – A Procedure for Conducting Conformity Assessment of AI Systems in Line with

the EU Artificial Intelligence Act. Available from https://www.sbs.ox.ac.uk/news/new-research-will-help-protect-society-unethical-ai.

[115] Partnership on AI. Human-AI Collaboration Trust Literature Review: Key Insights and Bibliography. Available from https://partnershiponai.org/paper/human-ai-collaboration-trust-literature-review-key-insights-and-bibliography/.

[116] Poretschkin M, Schmitz A, Akila M, Adilova L, Becker D, Cremers AB, et al. Guideline for Trustworthy Artificial Intelligence – AI Assessment Catalog. arXiv preprint arXiv:2307.03681. 2023.

[117] Prem E. From ethical AI frameworks to tools: a review of approaches. AI and Ethics. 2023; 1–18.

[118] Probasco ES, Toney AS, Curlee KT. The Inigo Montoya Problem for Trustworthy AI. Center for Security and Emerging Technology. 2023.

[119] Rawal A, McCoy J, Rawat DB, Sadler BM, Amant RS. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Transactions on Artificial Intelligence. 2021; 3(6): 852–866.

[120] Rawat DB. Towards Neuro-Symbolic AI for Assured and Trustworthy Human-Autonomy Teaming. In 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2023. pp. 177–179.

[121] Ronanki K, Cabrero-Daniel B, Horkoff J, Berger C. RE-centric Recommendations for the Development of Trustworthy (er) Autonomous Systems, TAS '23: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. 2023; 1–8.

[122] SAS. A Comprehensive Approach to Trustworthy AI Governance. Available from https://www.sas.com/sas/whitepapers/a-comprehensive-approach-to-trustworthy-ai-governance.html.

[123] SAS. Available from https://www.sas.com/en_us/company-information/innovation/responsible-innovation.html.

[124] Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M. Principles to practices for responsible AI: closing the gap. arXiv preprint arXiv:2006.04707. 2020.

[125] Shahrdar S, Menezes L, Nojoumian M. A survey on trust in autonomous systems. In Intelligent Computing: Proceedings of the 2018 Computing Conference. Vol. 2, 2019. pp. 368–386.

[126] Sheth A, Roy K, Gaur M. Neurosymbolic artificial intelligence (why, what, and how). IEEE Intelligent Systems. 2023; 38(3): 56–62.

[127] Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Interaction. 2020; 36(6): 495–504.

[128] Shneiderman B. Human-centered artificial intelligence: Three fresh ideas. AIS Transactions on Human-Computer Interaction. 2020; 12(3): 109–124.

[129] Shneiderman B. ACM TechBrief: Safer Algorithmic Systems, Association for Computing Machinery. 2023.

[130] Singapore's Approach to AI Governance. Available from https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework.

[131] Sousa S, Lamas D, Cravino J, Martins P. Human-centered trustworthy framework: A human-computer interaction perspective. Computer. 2024; 57(3): 46–58.

[132] Stettinger G, Weissensteiner P, Khastgir S. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access. 2024; 22718–22745.

[133] TAILOR network. Available from https://tailor-network.eu.

[134] TensorFlow Responsible AI. Available from https://www.tensorflow.org/responsible_ai.

[135] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 30. 2023.

[136] Thelisson E, Verma H. Conformity assessment under the EU AI act general approach. AI and Ethics. 2024; 1–9.

[137] Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electronic Markets. 2021; 31: 447–464.

[138] Toreini E, Aitken M, Coopamootoo KP, Elliott K, Zelaya VG, Missier P, et. al. Technologies for trustworthy machine learning: A survey in a socio-technical context. arXiv preprint arXiv:2007.08911. 2020.

[139] Trustible. Turnkey solution to maximize trust & make governance easy. Available from https://www.trustible.ai/future-why-trustible.

[140] TTC Joint Roadmap for Trustworthy AI and Risk Management. Available from https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management.

[141] UK Department for Science, Innovation and Technology. Portfolio of AI Assurance Techniques. Available from https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques.

[142] United Nations Educational, Scientific and Cultural Organization. Recommendation on the Ethics of Artificial Intelligence. Available from https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[143] Voirin JL. Model-based system and architecture engineering with the arcadia method. Elsevier. 2017.

[144] Xia B, Lu Q, Zhu L, Lee SU, Liu Y, Xing Z. From Principles to Practice: An Accountability Metrics Catalogue for Managing AI Risks. arXiv preprint arXiv:2311.13158. 2023.

[145] Xu W. Toward human-centered AI: A perspective from human-computer interaction. Interactions. 2019; 26(4): 42–46.

[146] Zicari RV, John B, James B, Boris D, Timo E, Todor I, Georgios K, et al. Z-Inspection®: A process to assess trustworthy AI. IEEE Transactions on Technology and Society. 2021; 2: 83–97.

[147] Zicari RV, Julia A, Frédérick B, Megan C, Boris D, Eleanore H, Alessio G, et al. How to assess trustworthy AI in practice. arXiv preprint arXiv:2206.09887. 2022.