WILEY | Hindawi

Review Article

Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements

Laith Alzubaidi [1,2,3] **Aiman Al-Sabaawi,**[1] **Jinshuai Bai,**[1] **Ammar Dukhan,**[1]
Ahmed H. Alkenani,[1] **Ahmed Al-Asadi,**[4,5] **Haider A. Alwzwazy,**[4] **Mohamed Manoufali,**[6,7]
Mohammed A. Fadhel,[2] **A. S. Albahri,**[8,9] **Catarina Moreira,**[1] **Chun Ouyang,**[1]
Jinglan Zhang,[1] **Jose Santamaría,**[10] **Asma Salhi,**[2,3] **Freek Hollman,**[3] **Ashish Gupta,**[3,11]
Ye Duan,[12] **Timon Rabczuk,**[13] **Amin Abbosh,**[6] **and Yuantong Gu**[1,3]

[1]*Gardens Point Campus, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[2]*Akunah Company for Medical Technology, Brisbane, QLD 4120, Australia*
[3]*Queensland Unit for Advanced Shoulder Research (QUASR), Brisbane, QLD 4000, Australia*
[4]*Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211, USA*
[5]*Communication Engineering Department, University of Technology, Baghdad 10001, Iraq*
[6]*School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4067, Australia*
[7]*Space and Astronomy, CSIRIO, Kensington, WA 6151, Australia*
[8]*Faculty of Computing and Meta-Technology (FKMT), Universiti Pendidikan Sultan, Tanjung Malim 35900, Malaysia*
[9]*Department of Computer Technology Engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad 00964, Iraq*
[10]*Department of Computer Science, University of Jaén, Jaén 23071, Spain*
[11]*Greenslopes Private Hospital and Queensland University of Technology, Brisbane, QLD 4120, Australia*
[12]*School of Computing, Clemson University, Clemson 29631, SC, USA*
[13]*Institute of Structural Mechanics, Bauhaus-Universiät Weimar, Weimar 99423, Germany*

Correspondence should be addressed to Laith Alzubaidi; l.alzubaidi@qut.edu.au

Given the tremendous potential and influence of artificial intelligence (AI) and algorithmic decision-making (DM), these systems have found wide-ranging applications across diverse fields, including education, business, healthcare industries, government, and justice sectors. While AI and DM offer significant benefits, they also carry the risk of unfavourable outcomes for users and society. As a result, ensuring the safety, reliability, and trustworthiness of these systems becomes crucial. This article aims to provide a comprehensive review of the synergy between AI and DM, focussing on the importance of trustworthiness. The review addresses the following four key questions, guiding readers towards a deeper understanding of this topic: *(i) why do we need trustworthy AI? (ii) what are the requirements for trustworthy AI?* In line with this second question, the key requirements that establish the trustworthiness of these systems have been explained, including explainability, accountability, robustness, fairness, acceptance of AI, privacy, accuracy, reproducibility, and human agency, and oversight. *(iii) how can we have trustworthy data? and (iv) what are the priorities in terms of trustworthy requirements for challenging applications?* Regarding this last question, six different applications have been discussed, including trustworthy AI in education, environmental science, 5G-based IoT networks, robotics for architecture, engineering and construction, financial technology, and healthcare. The review emphasises the need to address trustworthiness in AI systems before their deployment in order to achieve the AI goal for good. An example is provided that demonstrates how trustworthy AI can be employed to eliminate bias in human resources management systems. The insights and recommendations presented in this paper will serve as a valuable guide for AI researchers seeking to achieve trustworthiness in their applications.

1. Introduction

Our daily lives have been profoundly transformed by artificial intelligence (AI) and algorithmic decision-making (DM). In this computing society, AI and algorithmic DM influence most of our daily tasks, either directly or indirectly. Nowadays, with the immense available data, advanced algorithms, and high computing power, these systems have become increasingly complex and efficient. Understanding the logic behind these systems can be rather challenging, which explains the critical need to assess them, especially in terms of trustworthiness and reliability. AI systems can be fairly biased due to their constraints, biases, and ethical issues [1–3].

More importantly, data with biases and errors often result in unfair AI systems. For example, in the case of the CalGang database, a highly biased crime dataset with errors was used to train the AI system to predict violent gang-related crimes, resulting in a flawed system. Another example involves the use of the recidivism algorithm to predict the probability of reoffending, resulting in biased DM against dark-skinned people in the US courts [4]. Using a specific algorithm, Amazon was exposed to make recruitment decisions that put candidates at risk of discrimination and gender equity [5]. In addition to data training, continuous data flow is also necessary for these AI systems. This implies the importance of data privacy and governance against malicious activities. One notable example involves the Equifax data breach, which put millions of users at risk of their personal data being illegally exposed [6]. Furthermore, it is not a straightforward process to comprehend the logic of algorithmic DM, given the complexity of the advanced algorithms behind AI systems. As a result, the systems are not fully accepted or trusted. Despite the application and advantages of AI systems for the health industry, their trustworthiness is not widely accepted, further highlighting the issues of accountability in shared decision-making [7, 8].

Addressing all of these concerns, various methods and guidelines have been proposed to ensure the trustworthiness, reliability, and security of AI systems.

The trustworthiness of AI has recently gained the global attention of governments, major organisations, and scientific communities. For example, the International Organisation for Standardisation (ISO), which focusses on technical, industrial, and commercial standardisation, explored the attributes of accountability, controllability, fairness, and transparency of AI and recommended various methods to ensure the trustworthiness of AI [9–11].

Additionally, the European Union (EU) proposed ethical guidelines for trustworthy AI to govern and facilitate the development and operations of these systems [12] and the enforcement of the General Data Protection Regulation (GDPR) that promotes individual rights to explain AI decisions [13]. There are also frameworks that can measure and increase user trust in AI systems (proposed by the National Institute of Standards and Technology, NIST) [14] or promote accountability and responsibility for AI use (proposed by the U.S. Government Accountability Office, GAO) [15]. Furthermore, there is a specific programme known as explainable artificial

intelligence (XAI), which was launched by the Defence Advanced Research Project Agency (DARPA) [16, 17]. The programme focusses on making AI systems explainable and trustworthy. Considering the efforts made by these major organisations to ensure the trustworthiness of AI systems, the importance of trustworthiness in ensuring the success and security of AI for users and society is evident [18].

According to Gartner, about 30% of AI-based digital products would need a trustworthy AI framework by 2025 [19]. Moreover, 86% of the users reported their intention to place their trust and loyalty in companies that adopt ethical AI principles. These examples reaffirm the critical need to develop trustworthy AI. Consequently, there are different methods for different phases of the AI lifecycle. Certain methods are proposed to establish a solid foundation of trusting AI requirements and expectations during the design phase. In contrast, other methods are recommended for different phases of data collection, security, and preprocessing to ensure data variation, security, and fairness. There are also the modelling phase, which involves specific methods that ensure the explainability and interpretability of AI, and the implementation and oversight phases, which involve auditing and testing methods that ensure accountability and reliability of AI. According to the EU [12], human involvement is imperative in the development of trustworthy AI. The concept of collaborative intelligence is also recommended, which refers to the collaboration of humans and machines in DM [20]. In total, these methods share a common goal: developing trustworthy AI. AI systems must behave as intended with no unfavourable implications for users and society. Recently, artificial intelligence has been widely used in several applications, including those that are in touch with our lives (see Figure 1) [22]. Therefore, it is critical to make AI trustworthy.

Generally, there are three main factors in building trustworthy AI (see Figure 2): (i) trust in data to ensure that the data are free of bias, accurate, high quality, and privacy-preserving; (ii) trust in the AI model to address issues regarding model explainability, robustness, accuracy, and bias; and (iii) trust in the process to verify the evaluation process, compliance, and consistency. More details of these three factors are presented in the latter sections.

In light of the above, this review looks at the various aspects of trustworthy AI, including four questions and four contributions in response to each question.

(1) Q1: *Why do we need trustworthy AI?*

(2) Contribution 1: the reasons for the need for trustworthy AI have been explained

(3) Q2: *What are the requirements for a trustworthy AI?*

(4) Contribution 2: the key requirements that establish the trustworthiness of AI systems have been explained, including explainability, accountability, fairness, robustness, acceptance of AI, privacy, accuracy, reproducibility, human agency, and oversight
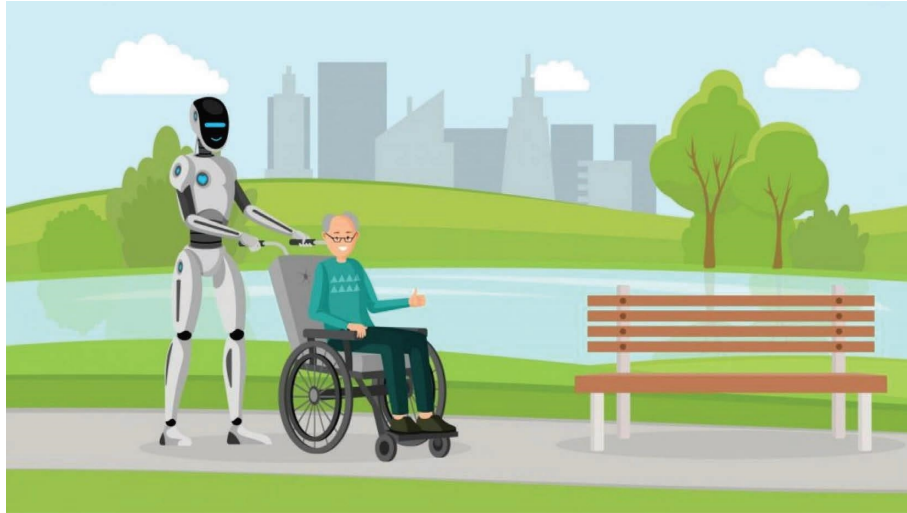
(5) Q3: *How can we have trustworthy data?*

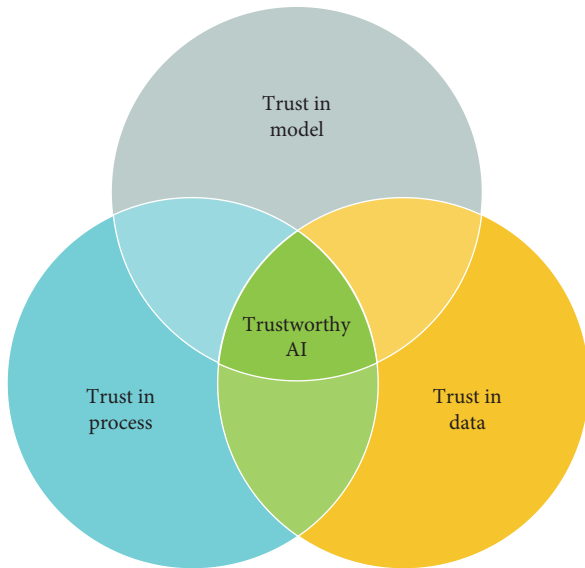FIGURE 1: Example of the future of robots which can be responsible for elderly people's care [21].

FIGURE 2: Main factors of trustworthy AI.

(6) Contribution 3: the challenges of the key requirements for preparing trusted data for AI, including high-quality data, privacy, and free of bias, have been explained.

(7) Q4: *What are the priorities in terms of trustworthy requirements for some applications?*

(8) Contribution 4: some guidelines are summarised for the prioritisation of different applications, including trustworthy AI in education, environmental science, 5G-based IoT networks, robotics for architecture, engineering and construction, financial technology, and healthcare

2. Search Strategy

We reviewed significant research papers in the field published during 2018–2023, particularly those from 2021, 2022, and 2023. Our comprehensive search was mainly performed in Science Direct (SD), Scopus, IEEE Xplore, and Web of Science (WoS), including the popular publishers IEEE, Elsevier, MDPI, Nature, ACM, and Springer. Some papers have been chosen from ArXiv. The selection of articles was based on the trustworthiness of artificial intelligence (AI) in different areas. Keywords were selected based on the recommendations of experts in the field of AI to answer all questions. Therefore, keywords related to applying trustworthy components in AI were specified as ("Trustworthy") AND ("Artificial Intelligence"), ("Trustworthy Artificial Intelligence") AND ("explainability"), ("Trustworthy Artificial Intelligence") AND ("accountability"), ("Trustworthy Artificial Intelligence") AND ("fairness"), ("Trustworthy Artificial Intelligence") AND ("acceptance of AI"), ("Trustworthy Artificial Intelligence") AND ("privacy"), ("Trustworthy Artificial Intelligence") AND ("accuracy"), ("Trustworthy Artificial Intelligence") AND ("reproducibility"), ("Trustworthy Artificial Intelligence") AND ("education"), ("Trustworthy Artificial Intelligence") AND ("environmental science"), ("Trustworthy Artificial Intelligence") AND ("5G-based IoT networks"), ("Trustworthy Artificial Intelligence") AND ("robotics for architecture") AND ("engineering") AND ("construction"), ("Trustworthy Artificial Intelligence") AND ("healthcare"), ("Trustworthy Artificial Intelligence") AND ("financial technology"), ("Trustworthy Artificial Intelligence") AND ("deep learning"), and Figure 3 depicts the search strategy.

3. Comprehensive Examination of Scientific Mapping

Due to increased contributions and applied research, it has become more challenging to identify crucial evidence discovered in previous studies. Keeping up with the literature has proved challenging due to the constant flux of practical and theoretical contributions. PRISMA is a technique that a number of academic specialists have proposed to reorganise previous studies' results, summarise problems, and
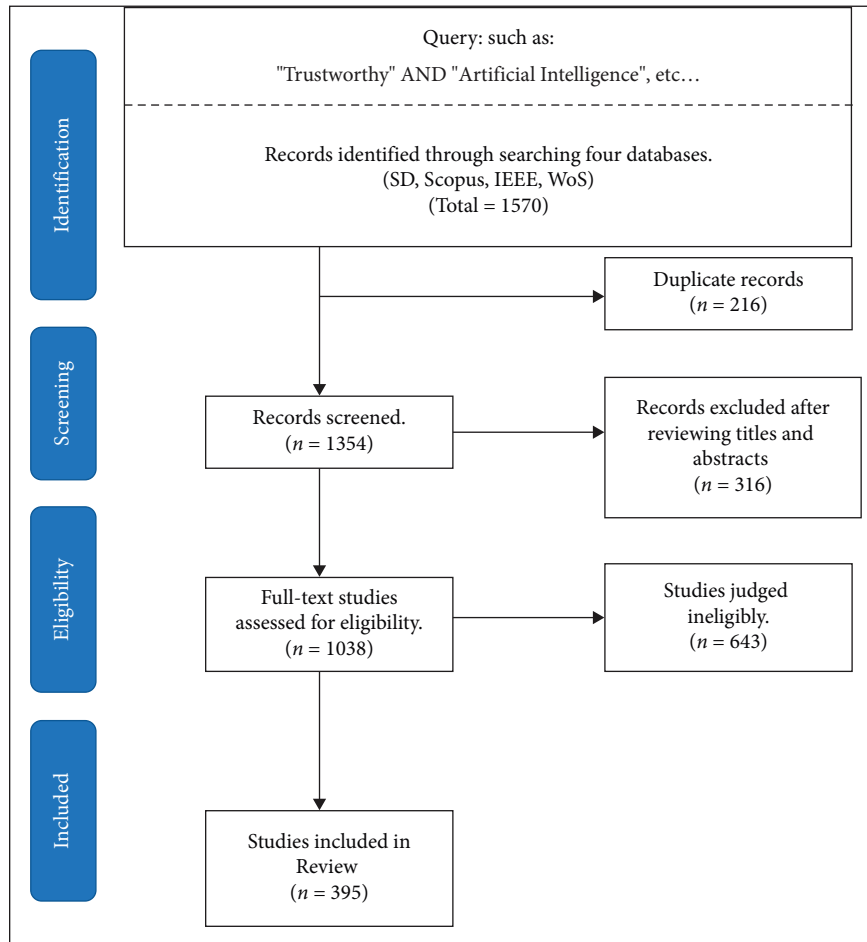
FIGURE 3: Search strategy.

identify future research gaps. In contrast, this paper presents a review that expands the scope of the knowledge base, strengthens the research strategy, and synthesises the literature's findings. However, despite their widespread use, evaluations continue to be beset by issues of reliability and objectivity. This is because such methodologies rely on the authors' perspective to reorganise the results of previous investigations. Several works have provided techniques for conducting an R-tool and a VOSviewer-based comprehensive scientific mapping analysis that is more suitable. The objective of these approaches is to promote openness in summarising the findings of previous investigations. The bibliometric technique yields conclusive results, identifies voids in the study, and draws conclusions about the findings of the corpus of literature in a highly dependable and transparent manner. In addition, the tools presented do not require high expertise and are considered open source. Consequently, the bibliometric methodology, which is described in detail in the following, was chosen for this investigation.

3.1. Annual Scientific Production. There has been significant progress in developing trustworthy artificial intelligence in the last 10 years. The yearly scientific output shown in Figure 4 explains the productivity of previous theoretical and practical investigations on reliable artificial intelligence. A limited number of articles were published during the first period from 2012 to 2015. Nevertheless, there has been a significant increase in published articles in recent years. The number of articles published increased markedly from 2017 to 2018, with a respective increase of eight to 27 articles per year. The number of articles saw a consistent upward trend in 2019 and 2020, culminating in a notable surge to 50 articles in 2020. The aforementioned trend remained consistent throughout 2021 and 2022, with 101 and 147 articles produced, respectively. The year 2023 is in its early stages, with a limited number of articles, namely eight, having been published so far. The overall pattern suggests a rise in the number of reputable research articles on artificial intelligence that have been published.

3.2. Three-Field Plot. The three-field plot is a common visualisation approach to show data with three separate parameters. The left field relates to sources (SO), the centre field to titles (TI_TM), and the right field to keywords (ID) in the specified situation. The graphic is often used to examine the correlations between the three parameters, as shown in Figure 5. According to the study
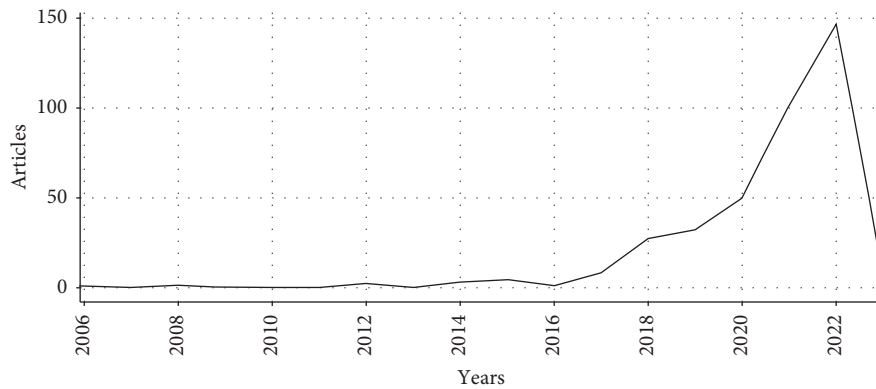
FIGURE 4: Annual scientific production.

title (TI_TM) in Figure 5, the Journals of Information Fusion, Automation Construction, and Nature have the most citations from the sources (SO) on the left side. Furthermore, the Information Fusion journal holds a significant place among the several publications that explicitly focus on the issue of dependable and understandable artificial intelligence (AI). Furthermore, the relevant domain (ID) acknowledges that the most frequently used phrases, such as "artificial intelligence," "trust," "robotics," "explainability," "information fusion," and "explainability," are consistently found in the journals indicated in the intermediate field (TI_TM).

3.3. Word Cloud. In prior studies, word cloud facilitated the identification of prevalent and significant phrases. This study aims to present and discuss critical findings from previous research, focussing on essential keywords. The purpose is to provide a concise summary and reorganise the material gathered. Figure 6 displays keywords of different widths. The considerable size of the keywords implies a higher frequency of their occurrence in the literature. In contrast, the diminutive dimensions of terms imply a lower frequency of occurrence. Based on the frequency distribution shown in Figure 6, prominent topics within the domain of trustworthy artificial intelligence (AI) include artificial intelligence, with an important emphasis on trustworthiness. Furthermore, the graphic illustrates that surveys and reviews are fundamental topics within the field. Topics such as ethics, privacy, and security often emerge, underscoring the need to consider these aspects throughout the development and deployment of AI systems. Figure 6 illustrates the many applications of artificial intelligence (AI) in several domains, such as healthcare care, decision-making processes, and information fusion. Additionally, this encompasses artificial intelligence methodologies such as convolutional neural networks (CNNs), natural language processing, and robotic systems. In general, the word cloud generated from trustworthy AI publications demonstrates the diverse nature of this field, including a broad range of topics. These include technical aspects of AI and its ethical, legal, and social implications.

3.4. Cooccurrence. A cooccurrence network is another tool used in bibliometric research. A semantic network is utilised in prior research to identify and examine widely used phrases, as well as in future analysis. This network provides professionals, policymakers, and researchers with vital insights into the underlying conceptual framework of a specific field of expertise. Figure 7 depicts essential information about a cooccurrence network built from the names of credible articles on trustworthy artificial intelligence (AI). The network is made up of nodes that represent the individual words in the titles. The edges of the nodes represent the frequency with which these words occur in the same title. Figure 7 depicts a set of nodes, as well as their clusters and proximity centrality values, which measure the degree of interconnection between a node and the other nodes in the network. The observation demonstrates that the nodes are divided into eight unique clusters, with the phrases within each cluster displaying a thematic or conceptual relationship with reliable artificial intelligence. Terms in Cluster 1 include "machine," "deep," "models," "trust," "adversarial," "privacy," security,' "detection," and "classification," suggesting a possible relationship with the development and implementation of reliable artificial intelligence systems. Artificial intelligence, trustworthiness, review, explainability, ethics, future, technology, opportunity, acceptability, health, industry, and responsibility are all words found in Cluster 2. These phrases suggest that the cluster is concerned with the ethical and explicable aspects of artificial intelligence. Similarly, additional clusters are associated with concepts such as "bias," "analysis," "fairness," "construction," "risk," and "safety." The closeness of a node in a network determines its centrality, which may be taken as a measure of its importance within the network. Words with higher proximity values have stronger links to other nodes in the network, indicating their greater importance with regard to the topic of trustworthy artificial intelligence. In general, the figure depicts the relationships between numerous concepts and phrases linked with the concept of dependable artificial intelligence, as evidenced in the titles of scientific works on this topic. The information provided may be useful in understanding the current level of research on this topic and in identifying areas that need more research.
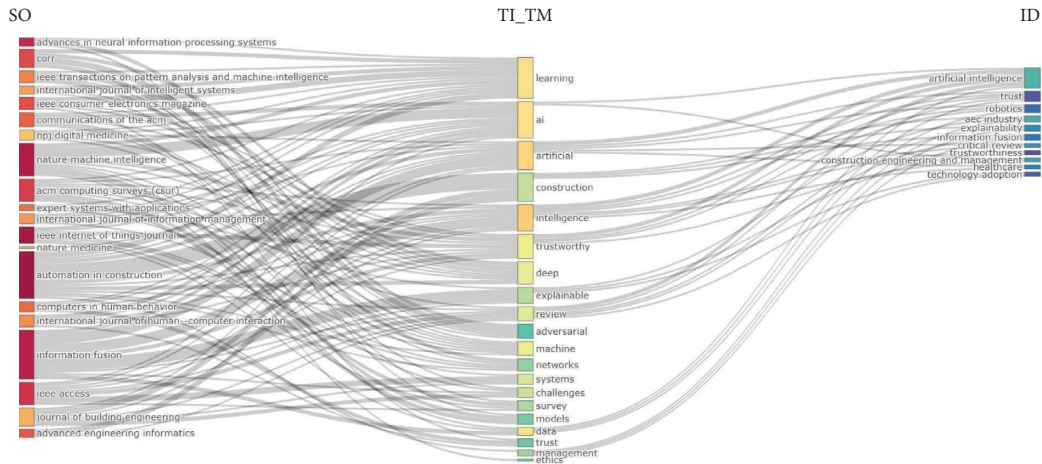
FIGURE 5: Three-field plot left (SO), the middle (TI_TM), and the right (ID).

FIGURE 6: Word cloud.
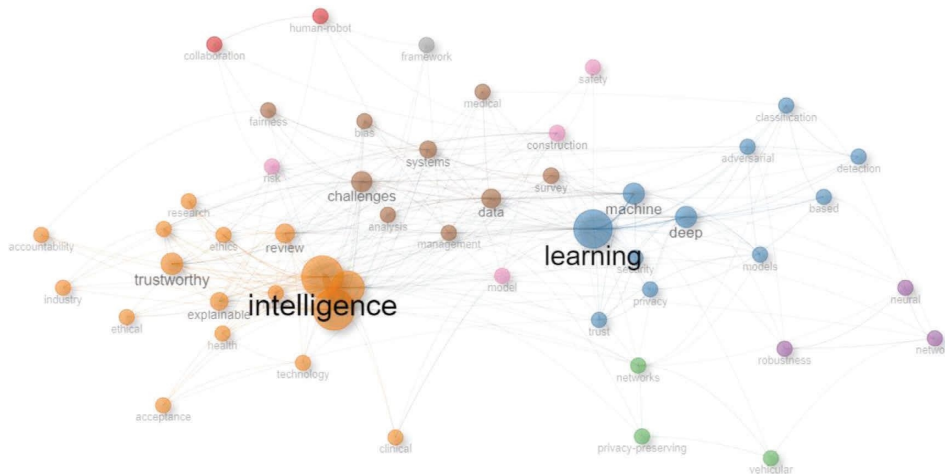
FIGURE 7: Cooccurrence network.

4. Related Work

In this section, we provide a comprehensive summary of the most recent evaluations and analyses conducted on articles related to the concept of trustworthy AI. These reviews represent a systematic effort to evaluate and critique the body of literature surrounding the development and deployment of AI systems with a focus on their trustworthiness, ethics, and reliability.

Nazir et al. [23] presented a survey of explainable DL methods for efficient and robust pattern recognition, categorising them into three main areas: explainable DL methods, efficient DL through model compression and acceleration, and robustness and stability in DL. The authors

of [24] examined the Trustworthy Artificial Intelligence (ALTAI) Assessment List, including its advantages and disadvantages. Additionally, they evaluated the tool's ability to assist the industry in understanding the potential risks associated with AI systems and implementing effective strategies to mitigate them. The importance of integrating research and methodologies from fields such as environmental sustainability, social justice, and corporate governance (ESG) is underscored in addressing analogous challenges in the ethical advancement and implementation of artificial intelligence (AI). Furthermore, this research investigates the prospective efficacy of the instrument with respect to its adoption within the sector, considering several factors that could impact its level of acceptability. The authors of [25] further guide professionals in the development of transparent artificial intelligence (AI) systems in healthcare. Moreover, their contributions serve to formalise the field of explainable artificial intelligence. The authors propose that the justification for seeking explainability is crucial in identifying the particular elements that need clarification. Consequently, this factor affects the relative importance of the characteristics related to explainability, including interpretability and integrity. The author introduced a theoretical framework to enhance the decision-making process by selecting several types of explainable artificial intelligence (AI) approaches. The categories mentioned above include the distinction between explainable modelling and post hoc explanation, as well as the many types of explanations, such as model-based, attribution-based, and example-based. Furthermore, the theory posits a distinction between global and local explanations.

The evaluation conducted by [26] offers a comprehensive examination of the latest developments in constructing an AI system with trustworthiness and explainability. This approach emphasises the intrinsic lack of transparency in artificial intelligence (AI), which limits our understanding of its underlying architecture. The article also explores several aspects of trustworthy AI, including its inherent biases and tendencies that undermine its dependability. This study explores the need to incorporate trustworthy AI into several industries, including banking, healthcare, autonomous systems, and the Internet of things (IoTs). This method enables the incorporation of trust-building tactics in several areas, including, but not limited to, data security, pricing, expenditure management, dependability, assurance, and decision-making processes. Trusted artificial intelligence is used across diverse domains and exhibits various degrees of adoption. Furthermore, it emphasises the need to use transparent and retrospective explanatory models throughout creating an explainable artificial intelligence (XAI) system. Additionally, this study outlines the potential constraints and obstacles that are likely to arise with respect to the development of explainable AI. This report provides a complete analysis of regulations related to the advancement of trustworthy AI in the autonomous car manufacturing industry. This paper explores several methodologies for developing artificial intelligence (AI) systems with reliability, interpretability, explainability, and trustworthiness. The overarching objective is to guarantee the safety of autonomous vehicle systems.

Numerous criteria, including fairness, explainability, accountability, dependability, and acceptability, have been proposed to increase the credibility of artificial intelligence systems. The survey [18] uses a methodology grounded in existing research to assess various requirements. It thoroughly analyses several techniques that try to mitigate the possible risks connected with artificial intelligence and improve the levels of trust and acceptance among users and society. The paper also discusses the analysis of existing approaches used in validating and verifying these systems, along with the continuous efforts in standardisation to guarantee the dependability and trustworthiness of artificial intelligence. The evaluation of confidence in artificial intelligence and robotics is of significant importance as these technologies continue to gain prominence within the architecture, engineering, and construction (AEC) sector. The publication [27] offers a complete review of the results obtained from an extensive analysis of the academic literature published over the last two decades. The study focusses on two main domains: (1) the notion of trust in relation to artificial intelligence (AI) and AI-driven robotics and (2) the many implementations of AI and robots in the architecture, engineering, and construction (AEC) sector. Also, analysis research thoroughly examines the shared characteristics of trust and assesses their applicability to current applications within the architecture, engineering, and construction (AEC) sector. The results indicate an increasing interest among scholars and professionals in the academic and industrial sectors of the architecture, engineering, and construction (AEC) domain with regard to investigating and using artificial intelligence (AI) and robotic technologies. However, extensive research that methodically analyses essential trust elements, such as explainability, dependability, robustness, performance, and safety, is still needed, particularly within the architecture, engineering, and construction (AEC) domain. On the contrary, the authors of [28] comprehensively analyse explainable and interpretable machine learning techniques in various healthcare fields. They also analysed the ethical implications of using ML and DL in healthcare. They also looked at the security, safety, and robustness problems that make ML less reliable. Furthermore, the study aims to comprehensively analyse how the incorporation of explainable and trustworthy machine learning (ML) methods might successfully tackle the ethical predicaments outlined earlier. The authors of [29] conducted a comprehensive and rigorous examination of previous findings, establishing a foundation for potential future investigations. The researchers thoroughly analysed the issues underlying the factors and proposed solutions related to the topic. The present study used a rigorous scientific mapping analysis to reorganise and integrate previous research results to resolve apprehensions about reliability and impartiality. Furthermore, the research has provided solid empirical data to substantiate the dependability of artificial intelligence (AI) in healthcare care. Addressing essential research gaps in this topic has been facilitated by presenting eight modern critical appraisals. The ubiquity of artificial intelligence (AI) in various aspects of human life has led AI systems to play a pivotal role in the

digital economy [30]. From the description, it seems that the map primarily focusses on the "trustworthy AI."

Table 1 presents a comparison between our survey and a more recent survey in terms of the requirements for trustworthiness and applications. Each of the articles mentioned focusses on a single or some requirement of trustworthiness, leaving a gap in the complete picture. This paper aims to present all the necessary requirements for trustworthiness, allowing researchers to have a comprehensive understanding and apply it to their AI applications. The primary goal of this paper is to synthesise and unify the various research efforts represented by the individual nodes on the map into a cohesive and interconnected understanding of trustworthy AI.

5. The Need for Trustworthy AI

Every person has a basic form of AI at home, for example, smart digital assistants, such as Siri, Alexa, and Google Assistant, robotic vacuums, or facial recognition [34]. This field has experienced revolutionary advancements, especially in the domains of machine learning and deep learning (DL), which were introduced in the early 2010s.

The tremendous prospects of AI have significantly expanded our imagination of reality with intelligent agents, contributing to prosperity and well-being at all levels. Despite that, similar to other technologies, AI systems come with ethical, legal, and social concerns [35]. Recognising the challenges of AI, there have been calls for "beneficial AI" [36], "responsible AI" [37], or "ethical AI" [38] in recent years. Various terminologies describe AI, but these terms reflect the same purpose, namely, to advance AI by maximising its benefits and minimising or preventing its adverse risks.

In early 2019, an independent organisation, specifically known as the high-level expert group on Artificial Intelligence of the European Commission, published the "Ethical Guidelines for Trustworthy AI" [39], which has attracted a great deal of interest from researchers and practitioners. These guidelines have also established a solid base for the use of the term "trustworthy AI" in existing guidelines and frameworks, such as the AI principles of the White House [40]. In particular, trust serves as a pivotal foundation for the economy of society and the country and sustainable development. The full potential of AI systems can be reached through a trusted AI foundation [41].

Defining and realising trustworthy AI is challenging. As a highly interdisciplinary and dynamic field, trustworthy AI covers various disciplines, ranging from computer science, economics, and management to sociology and psychology. There are varying views and understandings of what trustworthy AI is and different priorities on the ethical and regulatory criteria of trustworthy AI. Moreover, the technical and nontechnical aspects of trustworthy AI continue to evolve over time. As the complex phenomenon of trust alone has continued to spark scholarly debates in recent years, the conceptualisation of trustworthy AI and what constitutes trust in AI have remained inconclusive in both theory and practice. Currently, AI, especially DL, is involved in many applications such as education, environmental science, 5G-based IoT networks, robotics for architecture, engineering, and construction, banking, and healthcare [27, 42–44]. It is possible that the AI applications go wrong. Therefore, it is important to pay attention to how to make AI trustworthy. For example, in Australia, Robodebt, which is the government's automated welfare debt recovery process, led to the catastrophic failure to collect millions from welfare recipients [45]. Another example is that in July 2022, a robot broke the finger of a 7-year-old boy while playing chess, which is a red flag that reliability and safety are needed in AI [46]. There are more examples in which uncontrolled AI threatens enterprises (see Figure 8).

The beauty of AI is that if there is a bias, for example, then it can be fixed by adjusting the training data in terms of diversity, while it is hard to fix with humans. Therefore, it is critical to address these trustworthiness issues in the form of bias, hatred, and the propagation of bad ideas before transitioning from research to deployment. As a result, the goal is to have good AI systems capable of improving the quality of life.

6. Trustworthy AI Requirements

In this section, we will discuss the requirements for developing trustworthy AI applications (see Figure 9).

6.1. Explainability. When DM is more dependent on AI algorithms and systems, it is crucial that these decisions are explicable and can be trusted by different stakeholders. However, due to the inherent complexity of AI, it has become increasingly difficult to explain machine-learned decisions. Explainability should reflect the logic of the decisions made by AI systems and support system transparency and interpretability, leading to system improvement and better system governance [17, 33, 47–49].

An AI system that displays explainability facilitates the detection of errors or vulnerabilities and supports the establishment of the trustworthiness of the system [29, 50]. Users have the right to receive explanations about the results of an AI system, such as the thought process involved in the system to generate a particular result, the types of training data used by the system, and the metrics used to assess the validity and reliability of system results [51–53]. Additionally, an AI system should be able to provide appropriate explanations to users with different backgrounds, different expertise, and application requirements [54]. Users who clearly understand why an AI system produces a certain result are more likely to have a higher level of trust in the decision of the system [55]. It is important to understand that there are many forms of explanation that are driven by different purposes and to cater to different types of users, depending on their expertise and application requirements [56]. As a result, this has led to different levels of system interpretability, such as global interpretability and local interpretability [57, 58]. The methods of global interpretability focus on elucidating the entire logic and operation of an AI system. In other words, it provides an

TABLE 1: A comparison between our survey and a more recent survey in terms of the requirements for trustworthiness.

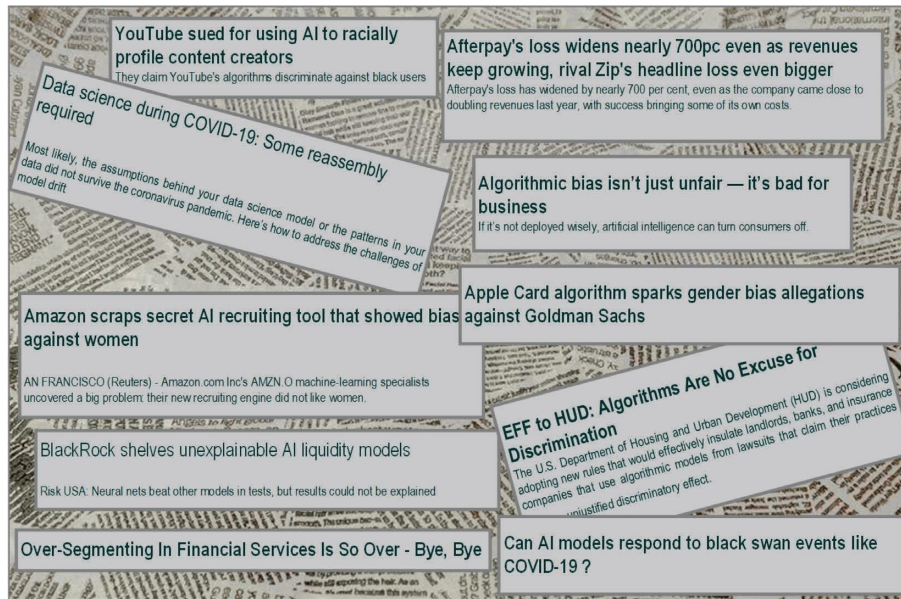| Ref	Application	Explainability	Accountability	Fairness	Robustness	Acceptance of AI	Privacy	Accuracy	Reproducibility	Human agency and oversight
Nazir et al. [23]	Biomedical imaging	Y	N	N	N	N	N	N	N	N
Ala-Pietilä et al. [24]	N	Y	Y	Y	Y	N	Y	Y	N	Y
Markus, et al. [25]	Healthcare	Y	N	N	N	N	N	N	N	N
Chamola et al. [26]	Banking, healthcare, autonomous system, and IoT	Y	N	N	N	N	N	N	N	N
Kaur et al. [18]	N	Y	Y	Y	Y	Y	Y	N	N	N
Emaminejad and Akhavian [27]	Robotics	Y	N	N	Y	N	Y	Y	N	N
Rasheed et al. [28]	Healthcare	Y	N	N	Y	N	Y	Y	N	N
Albahri, et al. [29]	Healthcare	Y	Y	Y	Y	N	Y	Y	N	Y
Vincent-Lancrin and van der Vlies [31]	Education	Y	Y	Y	Y	N	Y	N	N	Y
Feng et al. [32]	Speech-centric	N	N	Y	Y	N	Y	N	N	N
Chou et al. [33]	Healthcare	Y	N	N	Y	N	N	N	N	N
This survey	Education, environmental science, 5G-based IoT networks, robotics, finance, and healthcare	Y	Y	Y	Y	Y	Y	Y	Y	Y

Y: yes, N: no.

FIGURE 8: Uncontrolled AI threatens enterprises.

FIGURE 9: Trustworthy AI requirements.

explanations are provided after the decisions are revealed, validating the previous ex ante explanations [63]. As stipulated by ISO [9], ex ante explanations and ex post explanations are critical aspects of explainability for trustworthy AI in terms of transparency and interpretability.

Most of the methods to ensure explainability cater to the needs of system designers and developers for the purposes of debugging and oversight [64]. More appropriate methods are needed to support nonexpert users in dealing with the gap between transparency and actual implementation [65]. Additionally, organisations often have security and privacy concerns, resulting in a reluctance to adopt AI systems. Therefore, appropriate methods that take into account the explainability of AI systems should be developed without compromising users' privacy concerns and the security of these systems [66–69].

6.2. Accountability. How algorithmic DM is developed and operated must be monitored to avoid any unfavourable events or outcomes. These algorithms are simply computer programmes that are trained using data. Therefore, those involved in algorithmic DM are responsible for any adverse events or outcomes caused by the algorithms [70–72]. Wieringa [73] defined accountability as a networked account that assigns the responsibilities of stakeholders at different phases of the AI life cycle. Basically, algorithmic accountability involves evaluating algorithms based on relevant parameters and assigning responsibilities to those involved in the development of the algorithms.

The increasing use of algorithmic DM, especially in high-stake applications, has contributed to the critical need for accountability measures. The design, development, and application of these algorithmic models must be reliable and secure. System failure leads to irreversible losses and damages. For instance, the computer software of an aircraft had significant glitches, which contributed to the crash of
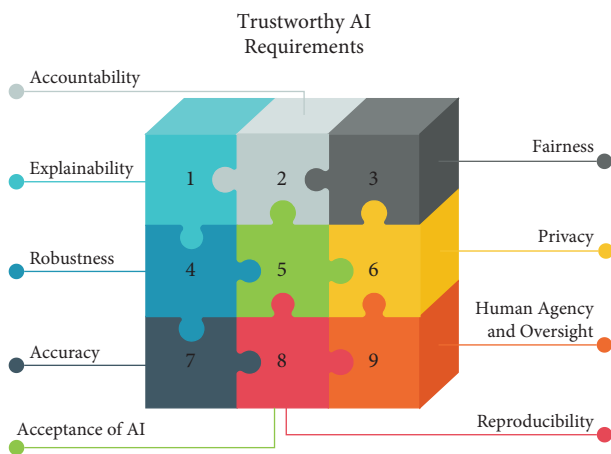
overview of the system in generating an explanation of the system's outputs and reasoning. This level of interpretability is used primarily to predict global trends, such as climate change [59], which can be rather challenging in terms of practise due to the scale involved. Meanwhile, local interpretability methods, which focus on elucidating specific decisions made by an AI system, are more commonly applied. The level of interpretability is more instantaneous than the global interpretability and distinguished according to when and for which input data the explanations are provided to the users [60, 61]. Accordingly, there are ex ante and ex post explanations under local interpretability [62]. With the purpose of establishing trust in AI systems, ex ante explanations elucidate the features, use, and operation of the systems prior to the actual use. These explanations imply that the systems are well-designed, tested, and validated. On the other hand, ex post explanations elucidate the features and conditions that result in the decisions. These

a Boeing plane and the lives of 346 people [74]. Volkswagen's electric car software had critical software architecture issues [75]. In another case, a face recognition system was found to be biased, putting women and dark-skinned people at a disadvantage [76]. All these failures can be prevented if only these algorithms are monitored accordingly. However, the question of who should be responsible for these failures remains: should it be the system developers, the ones who were in charge of data collection, or the system users who had the expertise to use the system? There is no conclusive answer to this question, which highlights the need to have a proper mechanism to establish accountability [77, 78].

There are different mechanisms to establish accountability of algorithmic DM, such as the incorporation of specific methods into the algorithm design process, the application of transparency methods, and the enforcement of strict laws and policies for better governance of algorithms. As an incremental process [79, 80], algorithms' accountability needs proper governance of the AI lifecycle and collaboration between different stakeholders [81]. Despite that, it is not a simple process to identify who should be responsible for system failure or to determine who exactly causes the said error since the development of algorithms involves different stakeholders. Appropriate accountability measures must be established based on the application domain since there is no universal measure applicable to all domains. For better governance, the development of context- and application-dependent accountability measures is proposed [9]. In line with that, ISO highlighted the use of medical AI systems and AI-based recruiting systems. In the case of medical AI systems, doctors who are the system users are responsible for any harm caused by the system, since they are the domain experts. More importantly, the system should only be used to help doctors in their DM process. Meanwhile, in the case of an AI-based recruiting system, the system users are not responsible for any harm caused by the system, since they are not in a position to know why their application is unsuccessful. This explains why context-dependent accountability measures should be further explored [82].

6.3. Fairness.

AI systems and algorithms rely on a tremendous amount of data and logic to execute a particular task and facilitate the DM process. Considering the substantial influence of AI systems in daily tasks and operations, it is important that these systems are not biased. A fair system does not discriminate against any individual or member of society [83]. The principle of fairness is in line with the concepts of ethics and moral values [84–88].

An unfair system in terms of its design, development, application, and monitoring is detrimental. There are several examples of unfair AI systems or algorithms. For instance, a judicial system was found to rely on an unfair risk assessment tool that discriminates against dark-skinned people [89]. In another case, a major tech firm was exposed to relying on an unfair hiring algorithm that discriminates against women [90]. Certain studies demonstrated discrimination against underprivileged

individuals or of certain races or ethnicities in the DM process of child maltreatment screening through the use of an unfair predictive analytic tool [91].

There are various factors that affect the trustworthiness of AI, such as data bias, model bias, and evaluation bias. With the critical need for fairness in AI, studies have proposed various definitions of fairness, but there is no conclusive definition for the term "fairness" in AI. Some previous studies compared and discussed these definitions in detail [85]. In general, how fairness in AI is defined depends on the context, specifically how the AI is applied. There are two broad definitions of fairness in AI: individual fairness and group fairness [85].

First, individual fairness makes sure that individuals of the same group receive similar predictions [92]. The definition of fairness under individual fairness is related to fairness through awareness [92] or unawareness [93], as well as counterfactual fairness [94, 95]. On the other hand, group fairness ensures that all groups of society receive equal treatment [96]. The definition of fairness under group fairness is linked to demographic parity [95], equalised odds [97], equal opportunity [98], and conditional statistical parity [99].

Apart from the various definitions of fairness, there are numerous methods available to establish fairness in AI. However, it is not a straightforward process to identify a single definition and method to detect all forms of bias. Feuerriegel et al. [100] recommended the need for more studies to explore the definitions and perceptions of fairness, particularly in relation to AI applications: certain AI applications may be more sensitive to certain elements than other applications. It is imperative to establish frameworks and policies that clearly define fairness based on the context of applications. Furthermore, different stakeholders view fairness differently, which suggests the need to involve various stakeholders to ensure the trustworthiness of AI. It is also important to have more robust testing methods and measures to identify and eliminate various forms of system bias [101].

6.4. Robustness.

According to the EU's ethical guidelines for trustworthy AI, robustness is one of the three essential criteria for building trustworthy AI systems [102]. IEEE defines the notion of robustness as "the degree to which a system or component can function correctly in the presence of invalid input or stressful environmental conditions [103]." This suggests that robust AI systems should be resilient against variations in input data or the external environment. For instance, in [104], the authors define the robustness of a deep neural network as "the integrity of the network under varying operating conditions, and the accuracy of its outputs in the presence/absence of input or network alterations," and propose that robustness is divided into two subproperties: security against possible attacks and reliability as resistance to environmental changes.

The majority of existing research on robustness in AI systems has focused on adversarial robustness, i.e., the ability to resist adversarial attacks [105]. Consequently, the term

(adversarial robustness) has often been used interchangeably with (robustness) in the literature. The purpose of an adversarial attack is to mislead an AI system into making an incorrect prediction by injecting into the system deliberately altered or manipulated input, known as adversarial examples [106]. These are a subset of perturbed examples and, unlike common perturbations, are generated to exploit weaknesses in the decision limits of the models [105]. For instance, adding a small perturbation to an input image results in a wrong decision (see the two examples [107, 108] as shown in Figure 10). It is critical to consider the issue of adversarial attack when it comes to building an AI system [109, 110].

Since adversarial examples can be used to test the robustness of an AI system, adversarial robustness is also considered the "worst-case robustness" [111, 112] and can be addressed as an optimisation problem [113]. In [114], the authors generate adversarial instances to attack a predictive model and assess the model's robustness by calculating the error rate of the model's prediction. In [111], the authors develop algorithms to search for the smallest additive distortions in the input space that are sufficient to confuse a classifier. In [109], the authors propose three major categories of robust optimisation methods, which are adversarial training to train predictive models using adversarial examples, certified defenses against norm-bounded adversarial examples, and regularisation approach to reduce the effects of perturbation on model predictions.

In recent years, nonadversarial robustness, which refers to "preserving model performance under naturally induced corruptions or alterations" in model inputs, has also drawn attention [115, 116]. Typical examples of corruption conditions include varying noise, blur, weather, and rendering conditions. In [117], the authors consider two main forms of nonadversarial robustness, known as corruption robustness and perturbation robustness, and propose benchmarks that evaluate model performance on common corruptions and perturbations for computer vision. In [118], the authors propose a set of corruption categories and specify a robustness score to measure the ratio of a model's corruption accuracy to its accuracy on clean input images. In speech recognition research, robustness to common corruptions (such as street noise, background chatter, and wind) is often more emphasized than adversarial audio since common corruptions are always present and remain unsolved [119].

In contrast to research on robust models for image and text data, robustness in AI systems that use structured data as input is underexplored. A few existing studies include detecting imbalanced and fraudulent tabular data using adversarial attacks to improve model robustness [120], adversarial training of predictive models that take time series data as input [121], and generating more event log data for predictive models using generative adversarial nets [122].

6.5. Acceptance of AI.

Trustworthiness is a must when it comes to AI, given the growing reliance on AI-based DM systems in our daily tasks. System failure leads to lower acceptance and a lack of trust in AI by system users. AI systems must be carefully evaluated using a specific mechanism in order to promote users' acceptance of AI and to gain trust in AI-based DM systems [123, 124]. Previous studies [125] identified several factors that help build trust, including system performance, type of task, type of application, human component and explainability of the system, and human involvement, which was identified as a crucial factor that can boost user confidence in AI systems and promote accountability for their decisions [125]. A few other prior studies [126] recommended having separate governance laws to promote the acceptance of AI. These recommendations share a similar purpose, to promote users' acceptance of AI through system evaluation. There have been efforts made to develop mechanisms to improve the acceptance of AI.

Furthermore, there are different expectations from different users of AI systems. Undoubtedly, the lack of information and the understanding of AI systems result in overstated or understated assumptions of systems in terms of system performance, usability, reliability, and fairness. Unfulfilled expectations lead to lower acceptance of AI systems. Some previous studies [127] demonstrated the low acceptance of AI systems by users despite the benefits of these systems and linked the mistrust of users in the systems with the lack of empathy and morality of the systems. Therefore, it is important to have a specific tool or mechanism for potential users to evaluate AI systems based on their trustworthy requirements, ethical principles, and expectations [128, 129].

6.6. Privacy.

AI-based DM systems involve a substantial amount of training data for their DM process. The amount of training data influences the accuracy and performance of these systems. However, there are implications to consider when it comes to the availability and application of data. The misuse of data by unscrupulous individuals, private companies, or governments has negative implications. For instance, a government misused the personal data of the citizens, resulting in inaccurate debt assessment [130]. In another case, the personal data of 50 million users of a social networking platform were collected and shared without their consent and misused to manipulate the US presidential elections [131]. These examples demonstrate the importance of ensuring data privacy, especially in gaining users' trust in systems [132].

Users are more likely to trust a system that applies the necessary mechanisms to protect their personal data and identity. The increased availability of data is linked to more cases of data breaches. For example, the Equifax data breach exposed the personal data of millions of users [133], and the details of credit and debit cards of 40 million Target customers were hacked [134]. Through such data breaches, a tremendous amount of sensitive information details are exposed and exploited. Internal attacks or targeted attacks lead to data breaches, resulting in system collapse and users' mistrust of AI. Therefore, securing the privacy of the AI systems and users
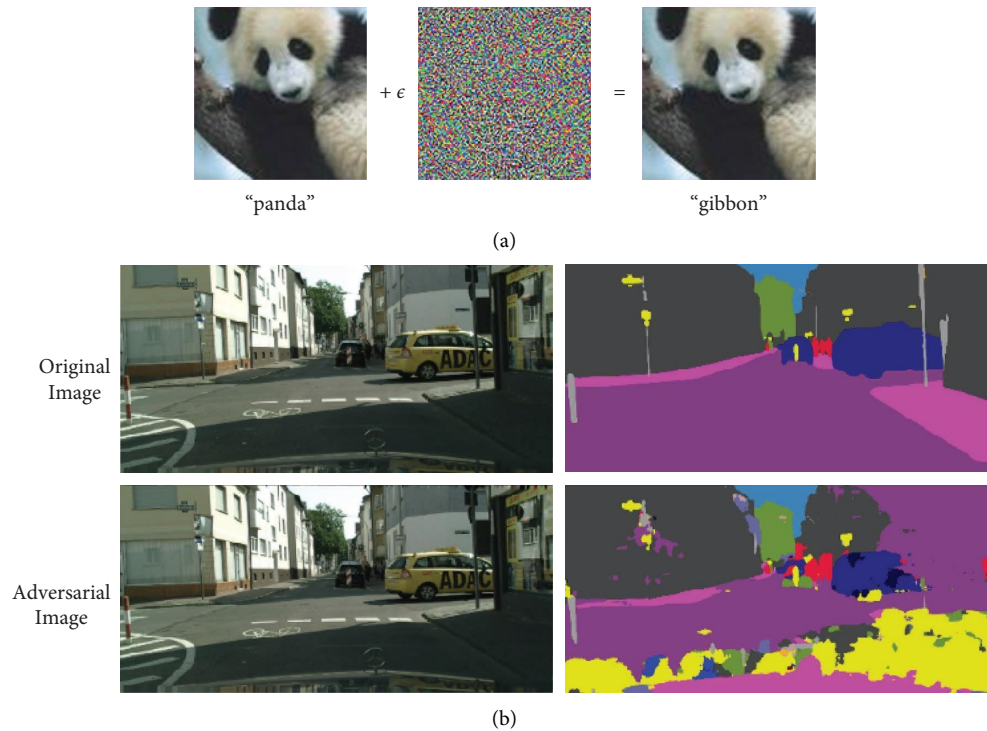
(a)

Original Image

Adversarial Image

(b)

FIGURE 10: Adversarial attacks illustrated by (a) a classification example [107] and (b) a segmentation example [108].

establishes trustworthy systems [135]. There are different phases (e.g., data collection, preprocessing, modelling, and implementation phases) of AI systems that are subjected to different privacy risks. For example, privacy risks associated with the data collection phase may lie in the gathered data or data storage. The inability to distinguish sensitive data from nonsensitive data during the preprocessing and modelling phases compromises privacy [136]. Frequently querying the model results in one's familiarisation with the internal functioning of these systems, which poses another form of threat that puts privacy at risk [137].

Studies have recommended various measures to enhance privacy, such as deidentification, federated learning techniques, and data agreements. These measures have their own strengths and limitations, depending on the requirements of the applications. How these measures are selected depends on the trade-offs between the performance and privacy-preserving overhead [138]. Despite the availability of these measures, AI systems are still susceptible to privacy risks. Appropriate laws are needed for AI systems to serve society and ensure system privacy, which can be challenging to accomplish both purposes at the same time. For instance, Rosenquist [139] highlighted the benefits of AI systems in rescuing exploited children. Still, the measure taken compromised the privacy of users—all images posted on the social media platform were subjected to facial recognition analysis, which is not possible for humans to perform. Considering that, the need for context-based privacy laws that clearly define the requirements and conditions that allow such measures to be taken is evident [140, 141].

6.7. Accuracy. Accuracy reflects the robustness of AI systems. A system that displays a certain level of accuracy yields reliable DM. An accurate AI system can deal with errors and problems and yields reproducible output with the same input under the example, a feature squeeze recommended various methods to ensure the input space-making items. For instance, a feature that squeezes problems reduces the complexity of the input space, making it less likely to be subjected to system problems [142]. Another proposed method involves incorporating problematic case examples into the system's training data [143, 144].

6.8. Reproducibility. With the same input parameters and conditions, a trustworthy system produces the same decisions, demonstrating its reproducibility [145, 146]. To quantify the state of measureity of empirical AI research, there are several reproducibility metrics measuring different degrees of reproducibility, including the hypotheses and how they were documented. Then, the method was explained in different aspects, including experiment, data, method, and new findings [147–149].

6.9. Human Agency and Oversight. Most importantly, AI systems must be subjected to the control of users as these systems continue to support and assist humans in their tasks [150, 151]. The control over these systems depends on the risks and implications of poor decisions on the users and society. The element of human agency and oversight ensures the involvement of humans in the DM process according to the risks, as well as social and environmental implications [152, 153].

7. Trustworthy Data for AI

One of the prominent issues in algorithmic DM lies in its data, specifically its availability and quality [154, 155]. However, the scope and potential use of data today are different. The significant amount of data contributes to the feasibility of AI [156]. An AI system can learn better and improve accuracy when supplemented with more data [157]. Advanced integration technologies are used to deal with interoperability limitations [158]. Open government data and the Internet of things are key sources of big data for governments and other institutions, such as data on education, health, public safety, social welfare, and taxes. Big data may not reach its potential for innovation without AI, suggesting their mutually enabling relationship. It should be noted that data with errors introduce risks to the systems, resulting in economic loss, mistrust in the systems, and compromised legitimacy. Therefore, data quality is key to establish trustworthy AI [155]. This section first discusses the challenges of data, including data quality and management, data security and privacy, and data bias. Then, we discuss measures to address these data challenges for trustworthy AI.

7.1. Large High-Quality Data. Making improvements to data analytics has successfully gained much attention, but improvements in data management for data analysis to establish trustworthy AI have remained underexplored. Furthermore, it is challenging to create a solid foundation for the development of AI systems, as no appropriate curated data resources are available [157]. Big data should not be viewed as data of better quality, and any data moved away from its original contexts do not reflect its original meaning [159].

There are two levels of the system that demonstrate the importance of data management: the operational level and the preprocessing level. Data management at the operational level enhances the overall trustworthiness of AI systems and confirms the availability of metadata to determine how data can be applied (e.g., data origin, format, extraction, organisation, classification, and connection) [160]. Meanwhile, data management at the preprocessing level confirms the reliability of data acquisition and the reproducibility of the results, confirming the transparency of the overall analysis.

It is crucial to train DL models using a large amount of data to achieve generalisation and ensure trustworthiness [161, 162]. However, data scarcity can pose a challenge in the training of DL models. To overcome this challenge, several techniques have been developed, including transfer learning [161, 163–167].

7.2. Data Security and Privacy. The significant potential of AI in simultaneously combining a large amount of data from multiple sources leaves trails of identifiable data, highlighting the significant relevance of privacy [168]. It is legally mandatory for government agencies and other institutions to protect and secure all personal and sensitive data; unfortunately, there are still cases of data breaches, such as the recent data breach at the Office of Personnel Management [169]. When combined and analysed together, integrated datasets provide the potential for "the ability to combine multiple customer views [that] may provide inappropriate insights" [170, 171]. The integration of multiple datasets contributes to the significance of AI and big data. It is a complicated phenomenon involving the privacy and application of data, in which no measures can increase data privacy without compromising the application of data [170].

7.3. Data Bias. AI models are math-based algorithms that themselves are not biased. However, the data can be biased. Therefore, it is critical to consider data diversity in terms of training AI models [172].

The bias in AI systems with respect to the data was classified into six groups [173]:

(1) Sample or selection bias when the database used underrepresents or overrepresents the target sample for the intended AI application

(2) Measurement bias relates to the systematic value misrepresentation when the instrument or the device used promotes a particular result

(3) Self-reporting bias is mainly present in the survey case, where absent, incomplete, and inconsistent responses might be provided

(4) Confirmation bias refers to observer bias when the working hypothesis is made based only on the observer's cognitive background and preferences

(5) Prejudice bias represents human data bias which reflects prejudice against gender, age, race, or ideologies, leading to discriminatory predictions and recommendations

(6) Algorithmic bias happens when the algorithm/model creates or amplifies the bias in an attempt to meet processing requirements

To address the challenges of data bias and make AI systems more trustworthy, we propose the following recommendations:

(1) Datasets used to train AI algorithms and build AI systems should be balanced and representative of the target population/estimations.

(2) Usage of different measurement devices/instruments, involving different observers, assessing the inter- and intra-observer reliability, and comparing the output, should be highly encouraged before engaging any dataset in AI pipelines. In fact, systematic measurement bias error cannot be addressed by simply collecting more data.

(3) Missing, incomplete, and inconsistent data should be disregarded.

(4) Whenever human observers are involved in the data curation of an AI pipeline, collection protocols and guidelines must be provided to ensure that the

relevant characteristics of the intended use are coherently represented. Furthermore, it should be comprehensive to take into account the opinions of different observers without misinterpretations.

(5) AI algorithm/model should be tailored to the available data. The algorithm design should reflect the intended use of the AI systems without creating or amplifying any data bias.

(6) Evaluate and benchmark the AI algorithms/models constructed from the balanced datasets. The best AI model should be selected based on multiple evaluation metrics simultaneously. Accordingly, the essential metrics weights must be considered when evaluating the developed mode.

7.4. Addressing Data Challenges for Trustworthy AI.

Addressing data challenges for trustworthy AI, studies have proposed various measures that governments and institutions should consider, which can generally be grouped into two forms of recommendations: data management and data literacy. First, data management measures serve as a fundamental foundation to establish trustworthy AI. The Data Management Association introduced the Data Management Body of Knowledge (DMBOK), which has become one of the most crucial data management frameworks [174]. It is an industry standard to evaluate data management and determine the appropriate strategies to manage data with the consideration of the following key dimensions of data management: (1) data governance; (2) data architecture and design; (3) database management; (4) data access (security) management; (5) data quality management; (6) master data management; (7) data warehouse and business intelligence management; (8) records management; (9) metadata management. Governance plays an important role in establishing the needed standards for all nine dimensions through specific structures and practices that exert authority and control (e.g., planning, monitoring, and enforcement) over the management of data assets. Through policies and standardised practices, institutions can systematically make data decisions and direct the functions of people and processes in data management [175]. Most importantly, data governance serves to manage data use with respect to established policies and practices toward achieving the key expected outcomes [176]. These sustainable and strategic data governance structures help minimise security issues and protect data creation, access, and application.

Second, data literacy plays an influential role in AI applications. Trustworthy DM relies on the integrity, security, and appropriateness of data in this digital era. Users, such as government employees, are responsible for the collection, analysis, and application of data. Through data literacy, users have knowledge of data management and have the ability to perform the tasks of creating, maintaining, and securing quality data to support daily operations [177]. Data cleaning, preparation, and review require meticulous documentation of judgment calls, further demonstrating the importance of data literacy [178]. These recommendations allow policymakers and decision-makers to evaluate and review the basis of AI development, promoting system transparency and accountability.

8. Trustworthy AI in Different Applications

This section presents six applications, including trustworthy AI in education, environmental science, 5G-based IoT networks, robotics for architecture, engineering and construction, financial technology, and healthcare. These applications are critical due to daily use; therefore, it is important to consider their trustworthiness in them (see Figure 11). The purpose of this section is to explain what the priorities are in terms of reliable requirements for the mentioned applications as listed in Table 2.

Building trustworthy AI systems is essential across various applications and industries. To achieve trustworthiness, it is crucial to prioritise requirements such as data privacy and security, explainability and interpretability, and adherence to regulatory standards. These requirements contribute to building trust, facilitating collaboration, ensuring ethical use, and improving outcomes in fields such as education, environmental science, 5G-based IoT networks, robotics for architecture, engineering and construction, financial technology, and healthcare. By addressing these requirements, AI researchers and practitioners can develop AI systems that are reliable, transparent, and effective, fostering widespread acceptance, and maximising the potential benefits of AI for society.

8.1. Trustworthy AI in Education.

There are various dimensions of trust in AI. There are different ways to deal with trust in AI for different countries. It is complicated to explain certain algorithms or techniques in layman's terms; therefore, a few countries like France have rejected the use of certain types of algorithms in public DM [179]. For China, it is about promoting social interaction and mutual trust. Meanwhile, the US take on expanding AI through research and development in order to improve the security, robustness, and trustworthiness of its systems closely mirrors the G20 AI Principles. Referring to the EU guidelines for trustworthy AI, there should be transparency, clear documentation, and identification of DM, and explainable technical processes and related human decisions. It is important that users of the AI system are informed of their interaction with the system and of the capabilities and limitations of the system [31].

Trustworthy AI performs what it is designed to execute in an unbiased manner. Taking the case of an AI-powered early warning system, students are subjected to profiling, and students who are at risk of dropping out are identified. An AI system that is not trustworthy may not be able to effectively identify these students or may be subject to misuse despite its accuracy in identifying these students. The system only matters when good (human) interventions are involved to support these identified cases of potential dropouts and deal with the potential risks involved. Addressing social justice, fairness, and nondiscrimination, in line with the G20 AI Principles, certain interventions may
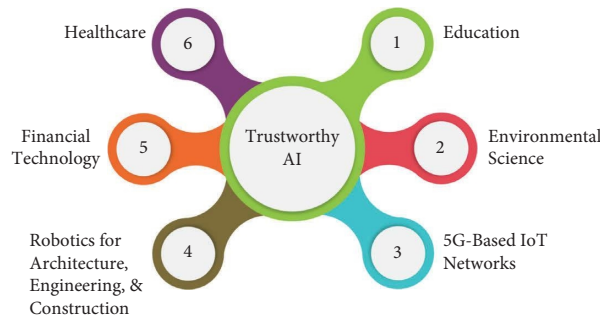
Figure 11: Trustworthy AI in different applications.

Table 2: Top three trustworthy requirements for each chosen application.

Application	Priority#1	Priority#2	Priority#3
Education	Privacy	Explainability	Fairness
Environmental science	Fairness	Explainability	Accuracy
5G-Based IoT networks	Privacy	Robustness	Accuracy
Robotics for architecture, engineering, and construction	Human agency and oversight	Robustness	Accuracy
Financial technology	Privacy	Accuracy	Explainability
Healthcare	Accuracy	Explainability	Fairness

serve to improve the school completion rate. Still, they may also exclude students who are at risk of dropping out due to certain accountability issues that potentially lead to unfavourable circumstances for the school (e.g., negative school reputation or image). In other words, the application of AI is not merely about establishing its trustworthiness but also involves a trustworthy connection between AI and humans [180].

The applications of AI in education are not widespread. However, there are various cases that can benefit from AI, such as automatic DM. For instance, AI can be applied for school, college, or university admission. This promotes fairness and unbiased selection of students. AI may introduce certain unintended implications since a new system is likely to alter the beneficiaries of the most in-demand schools, which is why it is important that the AI system and its algorithms display transparency and explainability. Transparency can be achieved by expanding "openness" to algorithms, but for certain AI techniques (e.g., deep learning), it is difficult to achieve explainability.

AI innovations benefit individuals and societies through education and learning outcomes, especially in this era of digitalisation. When it comes to trustworthy AI in education, privacy and security are among the most important aspects. Since students are generally minors, data protection and security are important issues, given how AI is largely based on the huge amount of personal data. The efficiency of AI in education increases with the use of personal data, but the collection and storage of these data pose privacy risks. Taking into account the fear of "Big Brother" in society, privacy concerns are magnified when it comes to AI in education. Large volumes of data are gathered and stored and can be retrieved at any point in time, raising concerns that "old" data are used in DM. There is also the possibility that data are exploited for commercial purposes. These

concerns reflect the "human-centered" nature of trustworthy AI in numerous ways. It is plausible that human-centred habits and practices are retained despite the prominence of AI: for example, the ability not to be constantly observed, not to risk having one's private information publicly exposed, and the ability not to be judged based on past or irrelevant information (but now available) [181, 182].

After all, each country has its own strategies to deal with privacy and security issues. For instance, when it comes to the European Union, the use of personal data is strictly regulated with respect to a stringent framework presented by GDPR, which emphasises transparency, data and storage limitation, and accountability. There are specific requirements for one to use, share, and store data. Similarly, the U.S. applies a specific framework on the use of personal data within the context of education under the U.S. Family Educational and Privacy Rights Act. Meanwhile, China implements the Chinese Governance Principles for Responsible AI, which promotes users' rights to be informed and to make their own selections freely and protects data privacy. In short, AI systems should consistently undergo necessary improvements in terms of controllability, explainability, transparency, and reliability without compromising the safety and security of these systems.

8.2. Trustworthy AI in Environmental Science. The applications of AI and machine learning (ML) in environmental science have recently gained growing popularity [183–186]. Environmental scientists employ AI/ML to make sense of raw data, e.g., satellite imagery and climate data, in order to come up with appropriate decisions for implementation. There have been more applications of AI, which have continued to improve predictions for various high-impact occurrences. However, unethically or irresponsibly developed AI may eventually cause more harm, which can be

observed in a few high-profile cases beyond the context of weather and climate [187, 188].

The benefits of AI in terms of environmental sustainability are undoubtedly evident, such as allowing automated monitoring of the ecosystems to support accountability for climate justice, for example, monitoring changes in land cover for the detection of deforestation [189–191], counting endangered species populations in very high-resolution satellite data [192, 193], and tracking bird populations in radar data [194–196]. Through automated analysis of retrospective and real-time datasets, relevant stakeholders can continue to monitor environmental trends and provide timely and effective responses.

Unfortunately, AI in environmental science may result in unfavourable results for certain application areas. Prior to the discussion on the shortcomings of AI in environmental science, the concept of environmental justice must first be discussed. Environmental justice refers to the unbiased treatment and meaningful public participation, regardless of demographic background and socio-economic status, to ensure the necessary development, implementation, and enforcement of environmental laws, regulations, and policies (EPA, 2022). With respect to this definition, there are multifaceted perspectives to take into account, from instrumental and consequential perspectives to principled ethical perspectives. These are entangled in the biases and pitfalls that this paper explores. Debiasing has the potential to address both. Note we assume the reader is generally familiar with the concepts of AI and ML and does not focus on any specific AI/ML methods but instead on the applications across environmental sciences. The question is "How AI can go wrong for environmental sciences?" The operation of AI depends on the quality of training data. Therefore, nonrepresentative or biased training data result in unreliable and biased AI systems and models. In other words, this phenomenon is known as "coded bias" [188]. For instance, an AI system that relies on training data that only contains hail occurrences in highly populated areas is more likely to show hail prediction in highly populated areas. Then, the system is biased to the population density aspect on which the system is trained. Algorithms that are frequently used in regions or circumstances that are different from where the training data are sourced may eventually perpetuate the bias beyond the regions and circumstances on which the algorithms are trained.

Issues related to training data are rather common and inevitable in most cases due to the following reasons: (1) it is extremely challenging to gather perfect, nonbiased datasets; (2) data developers themselves are not aware of the data limitations or biases; (3) the interpretation of bias depends on the context or application; (4) there are no standardised tests to check for common biases in datasets. Problematic data represent another common issue in AI [197, 198]. Such data may result in a faulty model or inaccurate model when real-world data are used in the model. When it comes to the quality of environmental science data, it is important to take note of the human and environmental factors that affect the data for use in AI. In addition to training data issues, there are also internal issues to consider. For instance, model

training choices influence every aspect of the AI model operation. Although the methods used to develop AI are generally "objective" according to specific mathematical equations, the following training choices can significantly influence the model outputs: (1) the selection of attributes for the model (e.g., environmental variables); (2) the source(s) of training data; (3) the selection of data preprocessing techniques (e.g., data normalisation, elimination of seasonality, and application of dimension reduction methods); (4) the selection of AI model type (e.g., clustering, random forest, or neural networks); and (5) the selection of hyperparameters (e.g., random forest in terms of the number of trees, maximal depth, or minimal leaf size). Each selection significantly influences the model outcomes, resulting in different results with critical implications. For example, the selection of spatial resolution is important to determine the model output for environmental justice—an AI model that is trained to predict urban heat at a low spatial resolution may overlook extreme values in small areas. In contrast, an AI model that is trained to predict urban heat at a higher spatial resolution can identify extreme values but potentially comes with noise.

On the one hand, faulty strategies are instead subjected to learning. AI models are trained to learn data patterns to produce good predictions. However, training data may not be representative of real-world settings. Faulty strategies can be identified by using interpretable or explainable AI (XAI) [199–202]. When it comes to interpretable AI, human experts must be able to understand the models [87]. However, explainable AI focusses on identifying the internal strategies used in more complicated models. Both interpretable and explainable AI enable human experts to evaluate the models for potentially faulty strategies prior to the application of the model [186].

8.3. Trustworthy AI in 5G-Based IoT Networks.

The rapid expansion of wireless broadband and multimedia applications in relation to the Internet of things (IoTs) has required enhanced capacity and robust quality of service (QoS). The reliance of IoTs on communication and computational structures potentially results in performance issues, as different sensors and devices are required for monitoring and control features [203–207]. The incorporation of 5G wireless networks can improve capacity and QoS in cellular networks, which can be subsequently addressed by addressing these bottlenecks. As a new decentralised architecture, the 5G network offers numerous benefits. Even at different locations, the resources for communication and computation can be acquired. The development of IoT-based vehicular communication technologies produces reliable wireless connectivity, enabling the development of intelligent transportation systems [208].

Accordingly, the 3rd Generation Partnership Project (3GPP) presented a roadmap that governs extremely high bandwidth, ultralow latency, and high-density connections to support 5G-empowered vehicle-to-everything (V2X) services. With 5G-enabled vehicular networks, vehicles with different sensors can be connected with each other through

mechanisms such as dedicated short-range communications (DSRC), LTE-V-Direct or device-to-device (D2D) millimetre wave, and by requesting and/or accessing resources from neighbouring vehicles, roadside units (RSU), 3GPP core network, or base stations [206, 209]. In addition to that, the prevalence of mobile and sensor-rich devices benefits IoT-based smart healthcare systems, particularly in supporting human activities, such as applications that recognise and monitor routine life-logging, healthcare, senior care, and personal fittings [206, 210, 211]. Similar to AI applications in other fields, IoT networks are also linked to security, privacy, and trust issues. Some academic and commercial institutions have made efforts to explore and develop vehicular networks to support this vision provided by 3GPP. Focussing on the issues of secure communication, studies have explored security, privacy, and trust in vehicular ad hoc networks (VANETs) [212–214]. Despite that, cybersecurity findings in the application of 5G in connected vehicles have remained scarce [213, 215–217]. Malina et al. [218] explored various IoT applications to categorise potential privacy leaks and then developed an innovative mechanism to protect privacy and maintain security within IoT services for actual applications, such as smart healthcare systems, intelligent vehicular networks, and smart cities. Using blockchain technology, Baza et al. [219] designed a privacy-preserving charging station-to-vehicle energy trading scheme, which can effectively identify Sybil attacks, protect the privacy of drivers, and significantly minimises communication and computation [220]. With the recent technological advancements in IoT networks, IoT users are able to make use of computational resources for IoT services. As modern vehicles in vehicular networks can exploit computational resources, drivers can have more relaxed, safer, and fuel-efficient rides, resulting in less traffic congestion and accidents. In such a scenario, the study of trustworthy privacy-preserving management in the computation is lacking and leads to the high potential risks to the vehicles [221, 222]. In particular, there are centralised and distributed mechanisms. For the centralised mechanism, data processing depends on the centralised servers. However, it does not fulfill the recent requirements, which is attributed to the immense number of users in large-scale networks (causes congestion at the centre). With 5G networks, systems can now fully make use of distributed nodes to address the significant computational burden in the centre, as well as the characteristics of IoT users and the unused computational capacity to process data in a distributed way. Meanwhile, as for the distributed mechanism, the trustworthiness of its management appears to be a major concern despite its benefits of improving the system's capability and reliability to complete computational service tasks. This may be attributed to the potential case of malicious miners deceiving the running algorithms, as well as the presence of egocentric computing nodes in the distributed computation (unwilling to share data) and the low processing capacity of miners, resulting in unfavourable contributions or even the divergence of algorithms. For both centralised and decentralised data processing, training data must be preprocessed or reconstructed to secure and protect any sensitive data, but this may reduce system accuracy. Therefore, robust privacy-preserving computation is necessary to compensate for the trade-off between system accuracy and data privacy [223, 224]. These issues must be addressed using a fair system design.

Recent advancements in AI techniques must be fully exploited [225–227] in the development of reliable privacy-preserving computations to improve data processing accuracy, protect sensitive data, and avoid adversary machine learning attacks [228]. Privacy is the most essential aspect that influences trust in AI when it comes to 5G-based IoT networks [229, 230]. With the recent emergence of new technologies and applications and their close connections with IoT networks, the privacy of IoT users has gained growing attention. Undeniably, technological advancements, especially in vehicular networks and smart cities (e.g., autonomous vehicles, hyper-connected vehicles, 5G and beyond networks, and AI-assisted Big data management in surveillance), have brought major improvements to the overall quality of life. However, security functions are brought to attention. For instance, an application that involves the capability to track trucks aims to identify the location of trucks and drivers in the terminals, maximise the throughput of container loading or unloading, and plan the driving routes according to the dynamic circumstances. With that, the application comes with an AI-assisted framework to perform the following functions: (1) mobility-aware attack detection; (2) realtime controller area network bus intrusion detection; (3) trust management system; and (4) privacy preservation of drivers. In other words, these functions help identify potential malicious actors in a triage manner without compromising trust, privacy, and security. In this case, the anonymity approach that takes advantage of the generation of pseudonyms as unique identifiers for authentication may benefit the privacy-preserving mechanism. Through such unique identifiers, any personally identifiable data are not required. Focussing on improving the conventional Hoepman's eight privacy design strategies, a previous study [218] demonstrated the successful application of such a privacy-preserving mechanism for private and secure IoT services, which included the preservation of data privacy (data collection, storage, and application), user authentication, communication, and computation. This simple yet effective design benefits numerous applications, such as identification systems, access control systems, smart safety systems, smart grids, and healthcare systems [231].

Apart from that, another approach to secure data communication involves the privacy preservation model under the data sanitisation process. In this sanitisation process, the key is optimally tuned by using many advanced AI-based algorithms, which include nature-inspired algorithms and DL techniques [205, 206, 232]. Moreover, the privacy-preserving model or the authentication mechanism can be designed according to the secure key management, which commonly involves secret key encryption or public key infrastructure with certificates [219, 220]. In the cryptography-based authentication process, the sender

generates codes using the secret key, and the receivers verify the codes attached to the message using a shared key.

Accordingly, there are two authentication methods: (i) authentication with central authority and (ii) authentication without central authority. For authentication with the central authority, an authority provides a unique identifier and a seed value (generating short-term pseudonyms), and the receivers maintain a random keyset for authentication while preserving their privacy [233]. However, this authentication method comes with communication and storage delays and limited nonrepudiation property. Meanwhile, as for authentication without central authority or, in other words, public key-based authentication, public or private key pairs are provided for pseudonymous communication, and the digital signature of a certification authority serves for authentication. Secret keys and short-term pseudonyms generate a digital signature, enclosed with its certificate in the data packet. Certification authorities are responsible for managing these certificate-based signatures, which are authenticated by receivers without exposing the identity of senders [234]. Considering the issues involved in the revocation of pseudonym certificates, efficient management of the certificate lifecycle is necessary. Taking the case of a vehicle certificate revoked due to fraudulence, error in certificate issuance, or compromised certificate, the driver cannot obtain new pseudonyms from the certification authorities, and it is challenging to authenticate pseudonyms against a certificate revocation list promptly due to the significant number of messages and lists [235].

In addition to that, there is signature-based authentication, which involves the use of identity-based signature, certificateless signature, or group signature [236, 237]. For this application, the private key is generated and assigned by the private key generator, and the public key represents the node identity for signature authentication. Meanwhile, the certificateless signature does not require certificates for authentication, and only a partial private key is provided by the key generation centre. It depends on the users to generate the actual private key according to its secret value and the partial private key or to generate the public key according to its secret value and the public parameters. When it comes to the group signature, the valid group members sign the messages anonymously. Although this approach maintains the privacy of vehicles, it is time-consuming to authenticate signatures. Therefore, it is not practical for real-time applications in VANETs [238–240].

In view of the above, studies on privacy preservation have placed emphasis on communication and overlooked its application in computation [241]. This highlights the critical need to establish a comprehensive privacy-preserving framework that deals with the required data communication and computation today, such as in the case of hyperconnected vehicles. Additionally, the integration of a privacy-preserving mechanism and trust management to create a unique design of privacy preservation and trustworthy management is clearly necessary. It is particularly challenging to design an effective trust model in 5G-based IoT networks. For instance, vehicular networks face inefficient long-term operations to connect vehicles due to the high mobility of users, resulting in constant or rapidly changing locations and topology updates, as well as numerous types of privacy and security attacks. This clearly calls for a trustworthy model that can simultaneously deal with these attacks and maintain data privacy [242]. To sum up, it is critical to establish a comprehensive privacy-preserving framework that deals with the required data communication and computation in 5G-based IoT networks.

8.4. Trustworthy AI in Robotics for Architecture, Engineering, and Construction.

For user acceptance, trust is a critical criterion for the interaction between technology and humans. Similar to other fields, the applications of AI in robotics for architecture, engineering, and construction have rapidly expanded, which calls for more studies on the trustworthiness of AI in these fields (see Figure 12), especially in terms of (1) privacy and security, (2) performance and robustness, and (3) reliability and safety [27, 244, 245]. The most crucial sociotechnical aspects of using AI and robotic technologies involve both privacy and security. Although privacy refers to one's right not to be observed, how AI operates requires extensive learning and continuous improvements involving humans [246]. Despite the linkage between privacy and security in AI, in most cases, both aspects should not be used interchangeably. In particular, privacy in AI involves the acquisition, analysis, and use of personal data. In contrast, security in AI involves protecting data confidentiality, maintaining data integrity, and ensuring prompt data availability upon request [247].

A hackproof AI system prevents data breaches, poor system design and engineering, unplanned data corruption, and malicious intent to obstruct or limit user access to the system [248]. Most recent studies on ethical AI tend to focus on the following topics [249, 250]: opacity of AI systems; privacy and surveillance; machine ethics or machine morality; the influence of ethical DM of automation on employment; manipulation of behaviour; the interaction between humans and robots; DM bias; autonomous system control; artificial moral agents; and singularity. For the interaction between humans and robots, privacy and security are relevant forms of risk to trust [27, 251]. The protection of cybersecurity systems promotes privacy and, subsequently, trust in robots [250].

Accordingly, there are two forms of risk to trust from the perspectives of privacy and security: (i) privacy situational risk and security situational risk and (ii) privacy relational risk and security relational risk. First, privacy situational risk refers to the belief that the system is likely to expose personal data involving users or their surroundings. Second, relational risk to privacy refers to the belief that the system is likely to expose users or their surroundings to unauthorised observation or disturbance. Third, security situational risk refers to the belief that the system causes users to be vulnerable to any form of threat to their safety. Lastly, relational security risk refers to the belief that the system is vulnerable to misuse, resulting in some form of threat to safety [251]. Ethical AI (responsible AI) promotes trustworthiness [27, 248]. To realise ethical AI, the security of data is

FIGURE 12: Example of the robotics for architecture, engineering, and construction [243].

required. However, only a few studies explored the ethical challenges of applying AI in robotics for architecture, engineering, and construction. Certain studies highlighted the failure of common worker monitoring methodologies (for the training of machine learning models) when considering workers' conscience, intentionality, and free will [252]. For applications involving AI and robotics, all data are gathered in monitoring construction tasks using telecommunications devices, wearable devices (e.g., VRs, smart, hard-hat cameras, and sensors), GPS, CCTV, drones, or smartphones [253–255]. Although there are different applications, ranging from worker safety [166] to equipment emission monitoring [256], data performance of workers and contextual information are indispensable for all applications.

Cloud computing is another example of a pervasive technology related to AI in the fields of architecture, engineering, and construction [257]. There are various privacy and security challenges of cloud computing in industries, which involve data security, access control, and intrusion prevention [258, 259]. Specific measures to improve data security in construction projects should be further explored [257] considering the importance of data security for such projects that are related to the military, government, and the public. These industries are often involved with confidential data, such as contract information, blueprints, images, and project personnel data. Careful attention is imperative to prevent any data breach [260].

Meanwhile, when it comes to the discussion on performance and robustness, machine competence in terms of technical performance and capacity serves as a key driver of trust in AI [261, 262]. The performance of AI is typically measured based on system accuracy. Accuracy functions as a performance benchmark to establish the trustworthiness of the system [261, 263]. A robust AI that is developed in a specific context can systematically deal with varying issues within and beyond the context system without compromising its performance [263]. From a theoretical point of view, performance and robustness may not go hand in hand; testing data issues can potentially result in an inaccurate classification at a high confidence level [264–266]. Despite that, both terms are typically considered when it comes to discussing the technical reliability of AI within the context of trustworthiness. In 2018, the European Commission prepared the Ethics Guidelines for Trustworthy AI. It defined "robustness" as "resilience, accuracy, trustworthiness of AI systems," which represents one of the seven key general trustworthiness requirements [22, 267]. Some of the proposed methods or techniques to promote robustness with acceptable accuracy include explicit training against known cases of system issues, as well as regularisation and robust inference [268, 269]. In most cases, users place their trust in AI based on their observation of its stated accuracy or the performance of the system in practice. When users observe that the accuracy of a system is low, their trust in the system declines, regardless of its stated accuracy [270, 271]. However, users would not trust the algorithm as soon as they identify any error, regardless of how the performance accuracy is observed [272, 273]. Most users establish their trust based on perceived accuracy [274], although trust is more affected by system failures than system successes [275].

Apart from system performance and robustness, there are other similarly effective performance metrics that contribute to trust building, such as precision and recall. For instance, users are more likely to emphasis the recall of classifiers instead of their precision in order to determine the acceptability of the performance of classifiers. However, the application can make a difference in terms of weight [276]. AEC use cases are not different from other applications in that the performance of the technology tool plays a key role in trusting and ultimately adopting it.

In construction, there is a specific target metric for the project cost or schedule [277, 278]. However, there are other comparatively important aspects to consider when adopting innovations and technologies, such as quality, safety records, sustainability measures, productivity metrics, and inspection results [279, 280]. Construction projects gain substantial benefits in terms of performance optimisation and assessment when BIM, big visual data, and modelling of construction performance analytics [281, 282] are adopted in conjunction with AI systems. The performance and efficiency of future projects can be significantly improved when the construction site layout planning process is digitised in BIM and training models. The development of AI-based real-time analytics tools and cloud-based data analytics for site data can facilitate projects to achieve performance targets and the required quality benchmarks [283]. For example, an AI chatbox can be used to receive, review, and share new projects and activities. After all, given the project-based nature of the construction industry, various contractors and other stakeholders with specialty trades or expertise are simultaneously involved. Therefore, AI applications, systems, and algorithms can be significantly different within a single project or from one project to another [284]. This is where the robustness of AI in construction comes into play. A robust AI must be transferable across varying projects, sectors, industries, and even geographic locations, given the uniqueness of any given project, without compromising user trust. For example, robots were used to pour concrete, brush uneven layers, and dismantle forms for a 334-metre-wide dam construction work [27]. The applications of AI also benefit tunnel construction projects, including inspection, maintenance, and health monitoring tasks [285, 286]. Robust models with the programming of specific project risk, quality, and inspection criteria are important for the effective and efficient transfer and application of tasks in construction projects. Both technical construction and business processes must demonstrate performance and robustness to establish trustworthiness [287].

In addition to that, reliability and safety serve as other key components of trust in AI-powered systems for architecture, engineering, and construction. [288, 289]. From the performance point of view (instead of the ethical point of view), reliability and safety are associated with trust in AI [290]. A safe interaction with AI reflects reliability and trustworthiness [291, 292]. However, a high level of reliability can contribute to overtrust and complacency [293], which explains the need to calibrate trust within the human-robot interaction. In the case of overtrust, users believe that

the system mitigates risks and eventually accepts too much risk [294]. Through trust calibration, potential risks can be accurately identified [295]. The process helps users to be aware of and familiarise themselves with the system features, consistency, and failure modes in order to prevent over- and undertrust [295]. In the case of undertrust, trust calibration helps users build trust gradually through experience and iterative interactions [296].

Besides that, the amount of effort by users versus the autonomy of AI affects the trustworthiness of AI from the reliability and safety points of view. As a human-robot interaction (HRI) factor, reliability is evaluated as a function of such autonomy. Beer et al. [297] proposed different classifications of robot autonomy (LORA) for HRI [297–299], starting from the level when users have full control over the process (manual) to the final level when the system has full control over the process (complete autonomy). These different classifications are reviewed according to the interactions between humans and robots or the evaluation of the sensing, planning, and acting tasks. In addition to LORA, there are classic Sheridan–Verplank levels of automation and recent self-driving car classification schemes [300, 301]. However, these taxonomies only consider circumstances when automation limits human control or, in other words, single-dimensional automation. Addressing these limitations, Shneiderman recommended the human-centred artificial intelligence (HCAI) framework to generate reliable, secured, and trustworthy designs [302]. According to HCAI, it is possible to simultaneously attain high levels of human control and automation (two-dimensional HCAI), which helps in enhancing the overall performance in terms of reliability, security, and trustworthiness [298, 302].

Numerous AI applications have been developed to address significant safety and health issues in the construction industry [303]. Furthermore, the use of AI to address safety issues in the construction industry has gained growing popularity [244, 304–306]. AI-controlled systems or robots can improve efficiency and minimise accident risks, which are highly favourable for high-altitude construction tasks, over a long period of time, or involve dangerous circumstances [307]. Common applications of AI for construction safety involve systems that make use of proximity monitoring of hazards. Taking the case of evaluating the interactions between construction workers and equipment, studies have proposed a system that can make use of the 5G wireless network to send images from trucks, cranes, and other construction machinery to a database for review [308, 309]. In addition to that, AI can be applied to predict appropriate safety measures, such as the severity of the injury, the type of injury, the affected body part, and the type of incident in a construction project [310]. For building construction, dynamic building information modelling (BIM) offers important information to the relevant stakeholders to develop more efficient planning, design, construction, and operation or maintenance. The use of BIM and AI-based software packages that incorporate machine learning algorithms can further improve the analysis of all aspects of a design, ensuring its reliability without

compromising other building systems [283]. For instance, Augusto and Nugent [311] extensively explored the benefits of AI in designing smart homes [312].

Despite the benefits of applying AI, its new applications may gain lower trust, given their potential safety risks. Studies have demonstrated the need for any new technologies or systems to demonstrate appropriate reliability and safety in construction [313]. When it comes to industrial applications of AI, human safety is a priority, since such applications are generally more robust and larger, which poses safety risks [314, 315]. Most previous studies on the reliability and safety of human-robot interaction in construction used two validation approaches. The first validation approach involves operating robot prototypes in physically simulated working environments. This serves to determine any weaknesses of robots in reliability and safety for enhanced functionality. The second validation approach involves developing predictive models to explain the perceived reliability and safety of robots and their effects on humans [316, 317]. For example, the robot acceptance safety model (RASM) integrates immersive virtual environments (IVEs) to review the perceived safety of collaboration between humans and robots. It involves several participants working together with a 3D simulated robot for a specific task using a head-mounted display. The results obtained revealed enhanced perceived safety when human and robot work areas are separated. This segregation promotes the identification of the team and the gradual trust in robots. In other words, the overall perceived safety and comfort of working with robots reflect their acceptance of future collaboration with robots [318, 319].

Any form of accident can affect the collaboration between humans and robots and the efficiency of work in industrial settings [320]. Addressing that, studies have proposed the use of active vision-based safety systems as one of the best measures. Various reviews and technical analyses have been conducted on the advancements of vision-based technologies and alarm systems in terms of methods, sensor types, safety functions, and static/dynamic actions of robots [321, 322]. Applications of AI and its subfields (e.g., machine learning, computer vision, knowledge-based systems, and natural language processing) and AI-powered robotics in construction mainly focus on safety and health to improve reliability and trust. However, the reliability and safety of AI in establishing cognitive trust are equally crucial and should be further explored.

8.5. *Trustworthy AI in Financial Technology.* With the development of new business models based on the utilisation of big data, fintech can confuse established financial intermediaries, especially banks [323]. Machine learning is a variant of AI that makes it possible for a computer to learn without an explicit programme. What is "deep learning"? It tries to derive meaning from big data using layers of learning algorithms. Costs may be reduced by applying new technologies, financial intermediaries, and improved consumer products [324]. AI-based applications need to be regulated to drive change in the right direction in compliance with global rules and standards especially when it comes to funds and sensitive data involvement. It is well known that trust is at the core of our financial system and building global trust in financial institutions is not easy. This is a bank regulation that works. AI-related policy recommendations and guidelines challenges for banks and other market participants (payments, third-party providers, Apps, and fintech) enforce strict control and create a secure environment so that the user is validated and identified [325]. With the combination of the right level of regulation and the ethical approach to artificial intelligence, ultimately user trust will be built and greater acceptance of AI-based practices in financial solutions.

To realise trustworthy AI in finance and insurance, the operational set of processes, methods, and tools is based on three challenges that are rarely addressed today.

(1) The model designer should choose the right topics of AI

(2) The model demonstrates valid accountability and discrimination, considering fairness criteria in data preparation and model design

(3) The model designer needs to achieve an unpleasant trade-off between accuracy and fairness or accountability

In this part, we will discuss the important concepts for addressing responsible AI in the financial sector.

8.5.1. *Explainability.* It determines how AI professionals and business leaders should address technology-specific limitations when explaining how AI models reach final and correct results. Despite the excellent performance of AI in a variety of areas, artificial intelligence (AI) is gaining in popularity. However, in addition to these achievements, lack of transparency, ambiguity, and the inability to explain and interpret most of the state-of-the-art technology are considered ethical issues [326]. Due to the complexity of AI, it is often difficult to explain and validate its predictions. This is sometimes referred to as the "black box" of machine learning [327]. Fintech is at the forefront of regulation covering accountability and equity issues. According to applicable regulations and guidelines, these systems need to be transparent. There are many different AI models today, ranging from linear regression to deep neural networks (DNNs). A simpler model may be easier to interpret, but its predictive power and accuracy are often poor compared to complex models [328]. These complex algorithms are very important in advanced fintech applications such as trading and cybersecurity. The models are intrinsically intractable models or approach interpretable models; these models are more complex and use reverse engineering [326, 327, 329, 330]. Implementing explainable fintech requires understanding the level of explainability according to the level of the user, for example, stakeholder, supplier, and third party. Thus, there are several points that should be considered when implementing XAI, including:

(1) Explanations should be familiar to the users and should be simple in terms of audience knowledge

(2) The XAI model should easily explain which features or variables influenced the model's predictions and at what step the decision is made

(3) The weaknesses and strengths of the model must be explained, as well as how it would work in the future

8.5.2. Privacy. One of the principles of responsible AI is privacy [331]. It is the ability to prevent people from obtaining information about others without their permission [332]. In recent years, the importance of data protection has received a great deal of attention. European Union (EU) and the Governor of California enact security laws and data protection policies to protect the rights to personal data. This gives more control to the consumers over their personal information. Therefore, the next generation of services while maintaining privacy is an open research topic [333]. Data protection is an important issue for companies that work with large amounts of data. When implementing AI solutions in the financial sector, the following items must be considered [334]:

(1) Data confidentiality and data subject privacy protection.

(2) Infrastructure security.

(3) Maintain the right to the level of personal access to the information. When collecting data for banking applications powered by AI solutions, the following guidelines should be established.

    (a) The personal information is collected and processed legally and fair.

    (b) The agencies should collect the right data to avoid harming the accuracy. Also, these data must be updated periodically.

    (c) The system that uses the data should apply security safeguards to avoid accidents.

The authors in [330] addressed two methods for preserving privacy in financial AI systems:

(i) Federated learning (collaborative learning) model: It allows multiple stakeholders who do not fully trust each other to work together to train the AI learning model on the combined dataset without data sharing. In this model, the local data are trained in local models, and the local models share the training parameters to complete the global model. This model can be centralised, decentralised, or heterogeneous.

(ii) Synthetic data-based model: It is artificially manufactured information that is not collected from the real world but can represent it. Using sophisticated AI algorithms, such as generative adversarial network (GAN) and real-world data, the synthetic data are generated. The newly generated information has similar statistical characteristics to real-world data. The newly generated data can be fully synthetic, partially synthetic, or hybrid synthetic [335],
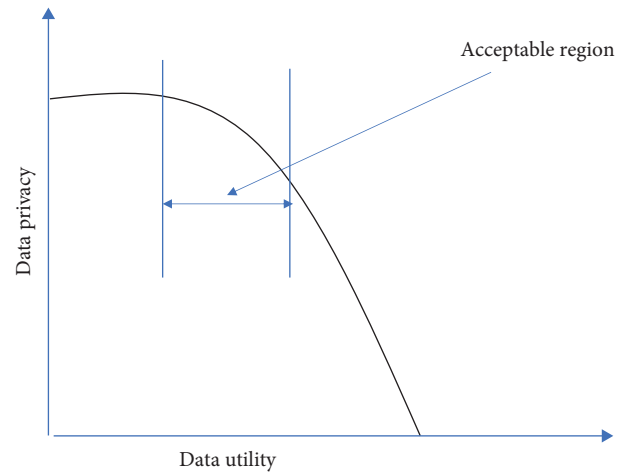

FIGURE 13: The trade-off between data privacy and data utility.

depending on the cost and privacy of real data. When using a synthetic dataset by the organisation, the utility metric should be evaluated because the utility indicates the validity and analytical completeness of the data; thus, a trade-off between utility and privacy perception should be considered. Synthetic data can be considered as a safeguard for personal information, which is critical in financial applications and important for anonymisation. Also, synthetic data are easy to access and exchange between organisations (see Figure 13).

8.5.3. Ethics. This is related to how AI treats society. All organisations must address this to build trust in their systems and meet the needs of their stakeholders for reliable information and accurate results [336]. The need for ethical AI systems is critical in the financial services sector. Finance organisations increasingly use AI to make critical decisions that can negatively impact customers, e.g., those related to credit card applications and rejection of loans. Therefore, to ensure ethical fintech, it is important to consider the following concepts when developing financial AI algorithms:

(1) The system should protect the privacy of the customers and their confidential data

(2) The system should be transparent

(3) The system should be unbiased

(4) The system should be accountable

AI can improve fintech, but there are limits. Specifically, bias can pose an ethical hazard that calls into question the reliability of the result generated by the system. Bias can be explained by the explainability of data, reproducibility when testing consistent results, and verifiability. Bias is not always bad and intentional. In some cases, it can generate discriminatory and unfair outcomes and is referred to as unfair prejudice. Bias can manifest itself in statistical models in many ways:

(1) Inherent in training data

(2) Algorithm distortion

In logic-based AI, bias is introduced depending on the rules that are affected by the knowledge of the designer. In a data-driven statistical model, the bias is affected by the way the data are collected. In bias free model, it is generated according to the manner in which the system is used.

8.6. Trustworthy AI in Healthcare. Similar to other industries, trustworthy AI is critical for the advancements and applications of AI in healthcare. It has become increasingly crucial to establish trustworthy AI systems following the growing dependence on AI in diverse healthcare applications. The ethics, evidence, bias, and equity aspects are key elements that contribute to the establishment of trust among healthcare professionals and patients [337–339]. The use of robotics was very helpful during the COVID-19 crisis and helped reduce the need for person-to-person contact [340] (see Figure 14).

There are fundamental values that ethical AI must maintain in healthcare care [338]. AI-enabled technologies in healthcare are different from the applications of AI in other fields—these applications in healthcare influence the therapeutic relationship between healthcare practitioners and patients. The training data for AI in healthcare provide varying quality and completeness and are intended to be applied in various settings and circumstances. This poses the risk of inequities in patient outcomes. However, AI systems are highly adaptive and have the capacity to learn and evolve over time beyond human observations and control [341]. However, the stakeholders involved with various forms of expertise, professionalism, and objectives are responsible for AI design, development, deployment, and oversight [342].

Despite the challenges of adopting AI in healthcare, there are frameworks for the ethical design and deployment of AI in this field. For example, the American Medical Association (AMA) Code of Medical Ethics on ethically sound innovation in medical practice (Opinion 1.2.11) specifies the need to establish the scientific foundation for any innovation that has direct consequences on patient care under the guidance of healthcare professionals with appropriate clinical experience, with minimal risks to patients and maximum benefits of the introduced innovation [343, 344]. Opinion 1.2.11 further elaborates on the need to have meaningful oversight in the development and integration of innovation into the delivery of care. According to the same code, Opinion 11.2.1 focusses on professionalism in healthcare systems, specifically on the ethical need to continuously monitor and report the results of AI innovations in the delivery of care [345]. In other words, innovations in healthcare should not put patients at a disadvantage or risk. They must be implemented based on the availability of resources and infrastructure for high-value care. Institutional oversight should also consider the possibility of adverse effects beyond clinical settings despite how well innovations are designed. For example, clinical prediction models are developed and adopted to identify individuals at risk of medical conditions but stigmatise or discriminate against certain individuals or communities. In 2019, the high-level expert group on Artificial Intelligence of the European Commission published the Ethics Guidelines for Trustworthy AI, which described the key role of trust in the development and adoption of AI and highlighted the need for a framework to achieve trustworthy AI [346, 347]. The publication described trustworthy AI as ethical, robust, and lawful. Its design should be human-centred according to ethical principles, including respect for human autonomy, prevention of harm, fairness, and explicability. The publication further noted the difficulty of observing and predicting the risks of AI systems, especially when vulnerable communities are involved, and the need for a holistic approach that involves multiple stakeholders and sociotechnical processes. On top of that, the publication saw AI as a sociotechnical system that needs to be reviewed within the context of the society for which it is designed [348].

Another recent study conducted by the European Parliamentary Research Service on "Artificial intelligence: From ethics to policy," conceptualised AI as a potentially beneficial and risky experiment in actual settings [348]. AI systems must be trained to be ethically responsible and to balance predicted benefits against potential risks without compromising the safety of humans. Recognising the highly promising prospects of AI, the study reported the significance of incorporating ethics into the design, development, and implementation of AI.

Apart from ethics, evidence plays an influential role in healthcare care, focussing on the validation of AI algorithms, but studies have demonstrated inconsistent terminology and approaches [349–351]. Achieving the highest scientific standards in design and development and providing clinical evidence of effectiveness and safety establish and enhance the trustworthiness of AI. There are specific frameworks for designing, conducting, and evaluating clinical research; for example, the US Food and Drug Administration (FDA) approved a drug and device development process that provides a model on which to base a standardised approach to meet this responsibility [352, 353]. An AI system designed for use in healthcare must at least demonstrate a clearly defined design protocol, address clinically relevant objectives and questions, and possess comprehensive documentation with scientifically rigorous and consistent validation in terms of safety and effectiveness. Besides that, a competent group of experts with diverse expertise and knowledge must review the AI system and prepare an unbiased report of its performance according to scientific standards.

These key requirements are subjected to continuous review and improvements as AI and technologies continue to evolve over time. Studies have proposed various methods to evaluate the quality and level of evidence required for the applications of AI in healthcare care. For example, the classification of recommendations, assessment, development, and evaluation (GRADE) evaluates the quality of evidence and the strength of suggestions from clinical practice [354]. In addition to that, there is a risk categorisation framework for software as a medical device (SaMD) developed by the International Medical Device Regulators Forum (IMDRF), which functions to categorise the impact
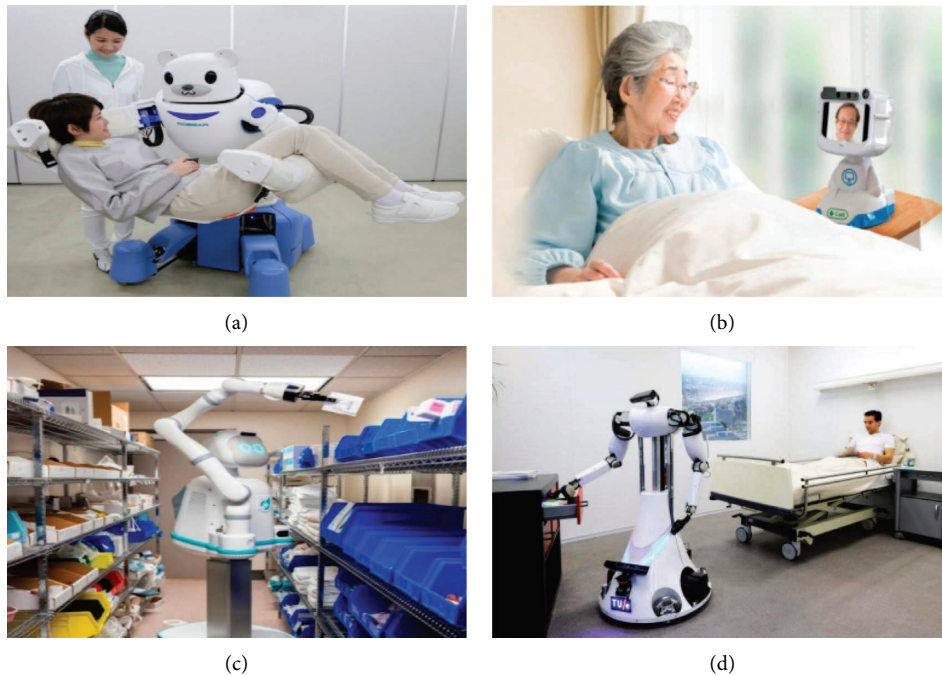
FIGURE 14: Nursing robots in hospitals and at home for elderly care. (a) Robear—a robotic bear nurse to lift patients in Japan. (b) Dinsow robot for elderly entertainment and face-to-face calls. (c) Moxi—nursing robot placing medicines in bins. (d) Robot attendant for hospital care (adopted from [340]).

on SaMDs based on the importance of the information provided for DM and the state of the healthcare condition [355, 356]. Such evidence and risk framework provide information on the levels of validation and evidence required for AI systems and deal with numerous ethical considerations, including sociotechnical and environmental considerations. Through the framework, IMDRF highlighted the importance of postmarket surveillance through a continuous learning process based on real-world evidence. Recognising the applications of AI in healthcare, ranging from administrative tasks to algorithms that inform diagnosis or treatment, acquiring evidence in proportion to the potential risks of AI for patients is highly crucial [357, 358].

Meanwhile, the applications of AI in healthcare care have highlighted the need to take into account the aspect of bias in the design, operation, or application of these adaptive systems in clinical settings [359–362]. Algorithms trained on electronic health records (EHRs), as most are currently, risk building into the model itself, whatever flaws exist in the record [363]. It should be noted that EHRs electronically capture information only from individuals with access to healthcare and that all data are not uniformly structured across these EHRs. Most of the data in EHRs consist of information captured "downstream" of human judgments. In other words, this comes with the risk that the model replicates human cognitive errors [363–365].

Despite all efforts to deal with potential biases in training data, this may result in unintended implications. For example, "race-corrected" algorithms divert resources from patients in the minority group instead of offering individualised and equitable care [350, 366, 367]. In addition, most

of the medical literature is Western medicine. Thus, this may not apply to different races, the developing world, and medicine in under-resourced regions. Thus, it auto-race corrected.

The development of unbiased models must address the definition of "fairness" [368, 369] and determine the appropriate trade-offs between fairness and performance [369, 370]. Furthermore, fairly designed algorithms, hypothetically, may become biased over time when these algorithms are applied in different contexts (from what they are designed in) or continuously trained on data with uncorrected biases in a broader healthcare system [359]. In certain cases, the designed models may be applied uncritically or within certain settings that are discriminatory or biased in nature. In addition to that, these models can have biased selection or a tendency to promote certain outcomes that are not in favour of the interests of individual patients [359, 371].

Focussing on health equity, the vision of AMA is to promote a vibrant environment that provides ample resources, equitable and safe systems, and the opportunity to achieve good health, as well as equips healthcare professionals with the awareness, equipment, and resources to deal with any inequities in all aspects of the systems. Despite significant technological advances and innovations to improve health equity, existing models of resource allocation, evidence development, solution design, and market selection have overlooked the incorporation of an equity lens, risking automation, scaling, and exacerbation of health disparities rooted in historical and contemporary racial and social injustices. The use of training data that excludes or under-

represents historically marginalised and minority groups contributes to the rise in equity. Significant differences in results related to the individual identity of the patients are not taken into account [372].

Furthermore, the design of the algorithm itself can exacerbate inequality when the related proxies or assumptions are discriminatory in nature. The nature of algorithms is generally more objective than that of humans, but these algorithms are developed by humans who are inherently biased [373, 374]. Marginalised communities are evidently under-represented in solution design and development, including those venture-backed start-ups, large technology firms, and academic medical centres. For instance, innovation teams and user testing efforts exclude any representations from the Black, Latinos, individuals with disabilities, and other populations. The Board of Trustees of AMA published a report on AI in healthcare back in 2018, identifying the hidden and unintentional existence of biases in training datasets that may be reproduced or normalised as a critical outcome for end users of AI systems impacts [347, 375]. According to a sociology professor and Princeton University professor Ruha Benjamin, Ph.D. in "Race After Technology," the book highlighted a similar notion of how "coded inequities" may appear neutral compared to historical discrimination but actually perpetuate and deepen discrimination and elaborated on the lack of intentionality as an inadequate attempt to rationalise the perpetuation of such discriminatory biases [376].

Regarding the European Commission statement on Artificial Intelligence, Robotics, and Autonomous Systems [377], the development and evaluation of AI solutions in healthcare must involve the intentional application of an equity lens from the start specifically from system design, development, testing, problem framing, training data selection, and algorithm design to the evaluation of the algorithm in order to identify, document, and eliminate these biases at the earliest stage possible. Although AI solutions are designed and developed more intentionally, AI is viewed as a downstream lever linked to larger upstream issues of inequity in the healthcare system. The application of these solutions remains bounded within a system that allocates the needed resources and opportunities for optimal health at the expense of others. However, healthcare professionals still have the opportunity to look upstream and explore beyond the design of the algorithm, including the health and care of patients under specific circumstances [368, 378]. Lastly, the issue with several state-of-the-art models is a lack of transparency and interoperability, which is a major drawback in many applications, including healthcare [379, 380].

There have been several tools to explain what is inside the "black box" of DL models and how the models make decisions [381].

Two examples have been listed from our ongoing work:

Example 1. Class activation mapping tool to visualise where exactly the DL model focusses on making a decision on a test set [382]. The shoulder classification task is chosen as an example (see Figure 15). Figure 16 shows the Grad-CAM and score Grad-CAM for shoulder X-ray images with two DL models. The first image shows the region of interest (ROI) in the red circle. The first model (image a) showed a correct prediction but could not be trusted because of low confidence and out-of-the-ROI focus. However, the second model (image b) showed a high correct prediction with a focus on ROI. It is necessary to investigate the results before moving to deployment.

Example 2. DeepDream is a technique that can be applied to display features that have been recovered by the network after the training phase [383]. In light of the fact that U-NET was shown to be a trustworthy segment in view of its strong statistics results, the resulting DeepDream images must set the primary features that differ per group (COVID-19 and non-COVID-19) (see Figure 17).

9. Challenges, Conclusions, and Future Directions

This paper discussed the significance and key requirements of trustworthy AI, which represents a fundamental research field in the AI ecosystem. The discussion began on the need for trustworthy AI, followed by trustworthy AI requirements, trustworthy data for AI, and lastly, applications of AI in education, environmental science, 5G-based IoT networks, robotics for architecture, engineering, and construction, financial technology, and healthcare. Accordingly, trustworthy AI demonstrates the following criteria: explainability, accountability, fairness, acceptance of AI, privacy, accuracy, reproducibility, and human agency and oversight. The development and working of trustworthy AI systems require appropriate measures, mechanisms, standards, and legal frameworks. Numerous studies have identified several technical challenges in developing trustworthy AI, such as the lack of clear requirements and standards to establish trustworthiness for AI. There are ambiguous definitions of the principles and properties of AI and unresolved differences in the principles across different application domains [9, 37, 384, 385]. For instance, a model that applies the principle of explicability is more prone to attacks, since the model is more interpretable and transparent. Therefore, stricter laws and trade-offs involving these principles according to the application domains are necessary.

Another challenge that affects the development of trustworthy AI lies in the need for context-specific solutions, as there is no universal solution for all problems. For instance, system developers may provide an explanation that is difficult for users with a nontechnical background to understand. A multidisciplinary team of experts is indispensable in the development of AI systems. In summary, trustworthy AI represents an emerging research field that requires more studies to further enhance the reliability and trustworthiness of AI.

Based on the findings of this article, there are several recommendations for future research on trustworthy AI. First, to optimise the potential of AI, especially in contexts that prioritise security, establishing the safety and trustworthiness of AI algorithms and systems is crucial. The
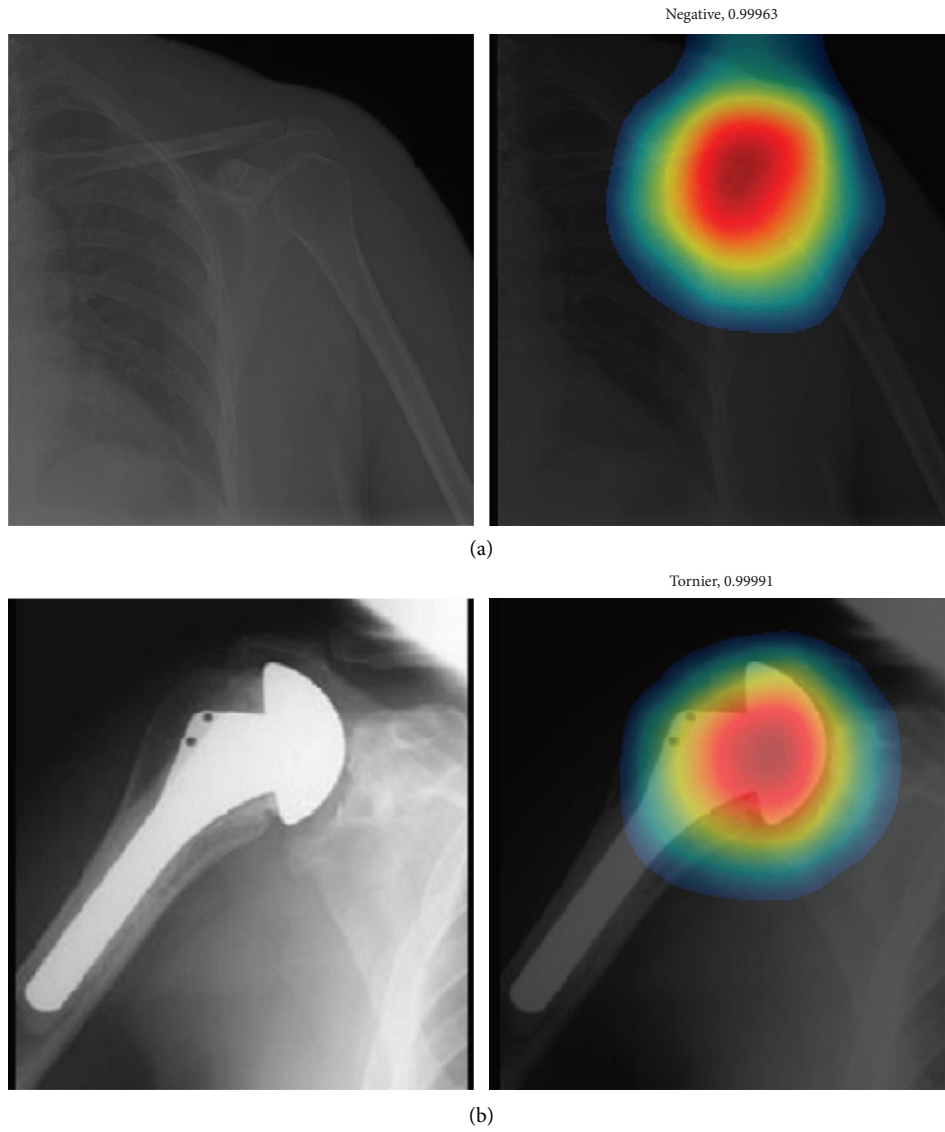
Negative, 0.99963

Tornier, 0.99991

(a)

(b)

FIGURE 15: Class activation mapping where (a) shoulder fracture classification into two classes negative and positive and (b) shoulder implant X-ray classification into four classes: cofield, depuy, tornier, and zimmer.

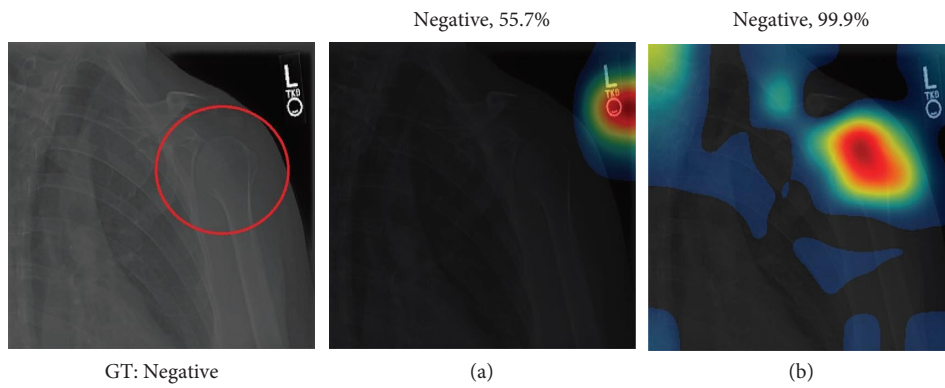Negative, 55.7%

Negative, 99.9%

GT: Negative

(a)

(b)

FIGURE 16: Grad-CAM and score Grad-CAM for shoulder X-ray image. The correct classification is negative. The first image is the original image with the label, and the red circle is the ROI. Images (a) and (b) are Grad-CAM and score Grad-CAM from two different deep-learning models.
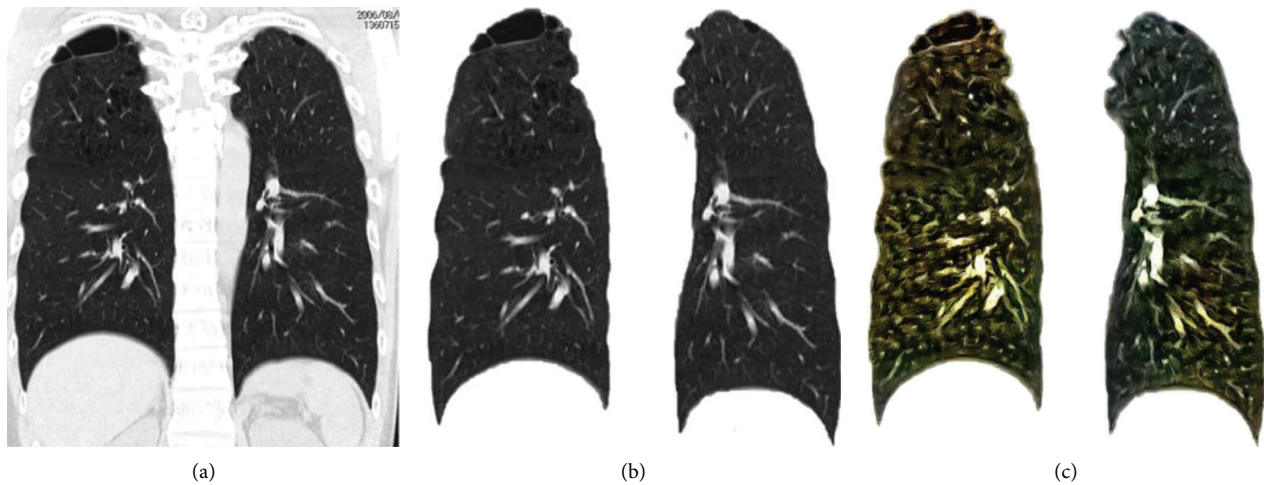
FIGURE 17: DeepDream where (a) original X-ray image of COVID-19 sample, (b) the end outcome of applying U-NET, and (c) DeepDream results emphasise the most important features.

ability to provide explanations to all stakeholders, ranging from system designers, developers, domain experts, and nonexpert users to policy makers, influences the trustworthiness of AI [386]. For that, multidisciplinary studies that involve computer science, data science, economics, law, and sociology are recommended to develop AI applications. This form of research offers significant knowledge and expertise from multifaceted perspectives that can enhance the safety and trustworthiness of AI.

Second, considering the growing importance of AI in this digital age, it has become increasingly crucial to establish standards and policies to ensure appropriate applications of AI. The standardisation of AI promotes more effective and efficient transfer of technologies, interoperability, security, and reliability. In addition to establishing new standards, AI applications must comply with existing laws and regulations according to their usage [18]. Therefore, future research is recommended to explore the needs and requirements of AI standards and enforcement policies.

Despite the significant benefits of AI, certain studies [127, 387] demonstrated the failure to use and accept AI among potential users, who reported the need for a certain mechanism that can objectively establish the trustworthiness of AI. Studies have proposed several theoretical models [388, 389], trust models [18], and human involvement methods [390, 391] to promote greater acceptance of AI. However, there are limited mechanisms that can measure and test user acceptance, which should be further explored. For example, future research is recommended to measure the effectiveness of trustworthy AI requirements and the implications of these requirements on user acceptance.

Last but not least, it is clear that there are different expectations about the capabilities of the system from different stakeholders, which influence the levels of acceptance and trust [392]. An expectation management framework from the beginning of the development of the AI system is crucial for system designers and developers to design and develop a system that is more well-received according to user expectations [302]. Further studies are also suggested on how to address postdevelopment expectation management to capture the influence of various factors such as system information, user reasoning and understanding of the system, and first-hand experience of the system in its acceptance by users [393–402].

Collaboration between AI experts and other experts in the domain, which is called "AI + X," is essential. It is important to bring them together to address any concerns and build a trustworthy AI. For the massive use of AI, it must be safe, trustworthy, and reliable. It is critical to ensure the development and deployment of trustworthy AI systems that benefit society while maintaining ethical and responsible practices.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

The authors contributed equally to this work.

Acknowledgments

References

[1] J. Zhang, M. Z. A. Bhuiyan, X. Yang, A. K. Singh, D. F. Hsu, and E. Luo, "Trustworthy target tracking with collaborative deep reinforcement learning in edgeai-aided iot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1301–1309, 2022.

[2] Y. Tai, B. Gao, Q. Li, Z. Yu, C. Zhu, and V. Chang, "Trustworthy and intelligent covid-19 diagnostic iomt through xr and deep-learning-based clinic data access," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15965–15976, 2021.

[3] X. He, Y. Chen, and L. Huang, "Toward a trustworthy classifier with deep CNN: uncertainty estimation meets hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[4] A. Eusebi, M. Vasek, E. Cockbain, and E. Mariconti, "The ethics of going deep: challenges in machine learning for sensitive security domains," in *Proceedings of the 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 533–537, IEEE, Genoa, Italy, June, 2022.

[5] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of Data and Analytics*, pp. 296–299, Auerbach Publications, Boca Raton, FL, USA, 2018.

[6] X. Fernández-Fuentes, T. Pena, and J. C. Cabaleiro, "Digital forensic analysis methodology for private browsing: firefox and chrome on linux as a case study," *Computers and Security*, vol. 115, Article ID 102626, 2022.

[7] E. Crigger, K. Reinbold, C. Hanson, A. Kao, K. Blake, and M. Irons, "Trustworthy augmented intelligence in health care," *Journal of Medical Systems*, vol. 46, no. 2, pp. 12–11, 2022.

[8] K. A. Crockett, L. Gerber, A. Latham, and E. Colyer, "Building trustworthy ai solutions: a case for practical solutions for small businesses," *IEEE Transactions on Artificial Intelligence*, vol. 4, p. 1, 2021.

[9] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, vol. 31, no. 2, pp. 447–464, 2021.

[10] N. Hasani, M. A. Morris, A. Rahmim et al., "Trustworthy artificial intelligence in medical imaging," *PET Clinics*, vol. 17, pp. 1–12, 2022.

[11] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799–819, 2023.

[12] E. Lemonne, *Ethics Guidelines For Trustworthy Ai*, Future European Commission, Brussels, Belgium, 2001.

[13] E. Hickman and M. Petrin, "Trustworthy ai and corporate governance: the eu's ethics guidelines for trustworthy artificial intelligence from a company law perspective," *European Business Organization Law Review*, vol. 22, no. 4, pp. 593–625, 2021.

[14] AIME Planning Team, *Artificial Intelligence Measurement and Evaluation at the National institute of Standards and Technology*, National Institute of Standards and Technology, Washington, DC, USA, 2021, https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop.

[15] D. Almeida, K. Shmarko, and E. Lomas, "The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of us, eu, and UK regulatory frameworks," *AI and Ethics*, vol. 2, no. 3, pp. 377–387, 2021.

[16] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, Article ID eaay7120, 2019.

[17] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022.

[18] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–38, 2022.

[19] B. Burke, D. Cearley, N. Jones et al., *Gartner Top 10 Strategic Technology Trends for 2020-smarter with Gartner*, 2021.

[20] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation," *Information Systems Research*, vol. 33, no. 2, pp. 678–696, 2022.

[21] E&T Editorial Staff, "Nursing care robots become more human with improved control method," 2020, https://eandt.theiet.org/content/articles/2020/01/nursing-care-robots-become-more-human-with-improved-control-method/.

[22] A. Holzinger, M. Dehmer, F. Emmert-Streib et al., "Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence," *Information Fusion*, vol. 79, pp. 263–278, 2022.

[23] S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, vol. 156, Article ID 106668, 2023.

[24] C. Radclyffe, M. Ribeiro, and R. H. Wortham, "The assessment list for trustworthy artificial intelligence: a review and recommendations," *Frontiers in Artificial Intelligence*, vol. 6, Article ID 1020592, 2023.

[25] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, Article ID 103655, 2021.

[26] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (xai)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[27] N. Emaminejad and R. Akhavian, "Trustworthy ai and robotics: implications for the aec industry," *Automation in Construction*, vol. 139, Article ID 104298, 2022.

[28] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: a survey," *Computers in Biology and Medicine*, vol. 149, Article ID 106043, 2022.

[29] A. Albahri, A. M. Duhaim, M. A. Fadhel et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion," *Information Fusion*, vol. 96, pp. 156–191, 2023.

[30] G. Li, B. Liu, and H. Zhang, "Quality attributes of trustworthy artificial intelligence in normative documents and secondary studies: a preliminary review," *Computer*, vol. 56, no. 4, pp. 28–37, 2023.

[31] S. Vincent-Lancrin and R. van der Vlies, *Trustworthy Artificial Intelligence (Ai) in Education: Promises and Challenges*, The Organization for Economic Cooperation and Development, Paris, France, 2020.

[32] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni, and S. Narayanan, "A review of speech-centric trustworthy machine learning: privacy, safety, and fairness," *APSIPA*

Transactions on Signal and Information Processing, vol. 12, no. 3, 2023.

[33] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022.

[34] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: on the past, present, and future of artificial intelligence," *California Management Review*, vol. 61, no. 4, pp. 5–14, 2019.

[35] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Machine Learning and the City: Applications in Architecture and Urban Design*, John Wiley & Sons, Hoboken, NJ, USApp. 535–545, 2022.

[36] S. Russell, "Provably beneficial artificial intelligence," in *Proceedings of the 27th International Conference on Intelligent User Interfaces*, p. 3, New York, NY, USA, March, 2022.

[37] P. Mikalef, K. Conboy, J. E. Lundström, and A. Popovič, "Thinking responsibly about responsible ai and 'the dark side'of ai," *European Journal of Information Systems*, vol. 31, 2022.

[38] M. Ashok, R. Madan, A. Joha, and U. Sivarajah, "Ethical framework for artificial intelligence and digital technologies," *International Journal of Information Management*, vol. 62, Article ID 102433, 2022.

[39] L. Floridi, "Establishing the rules for building trustworthy ai," *Nature Machine Intelligence*, vol. 1, no. 6, pp. 261-262, 2019.

[40] J. Chapiro, B. Allen, A. Abajian et al., "Proceedings from the society of interventional radiology foundation research consensus panel on artificial intelligence in interventional radiology: from code to bedside," *Journal of Vascular and Interventional Radiology*, vol. 33, 2022.

[41] I. Ulnicane, "Artificial intelligence in the European Union: policy, ethics and regulation," in *The Routledge Handbook of European Integrations*, Taylor & Francis, Oxfordshire, UK, 2022.

[42] S. K. Lo, Y. Liu, Q. Lu et al., "Toward trustworthy AI: blockchain-based architecture design for accountability and fairness of federated learning systems," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3276–3284, 2023.

[43] J. Ma, L. Schneider, S. Lapuschkin et al., "Towards trustworthy ai in dentistry," *Journal of Dental Research*, vol. 101, Article ID 00220345221106086, 2022.

[44] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, J. Zhang, J. Santamaría, and Y. Duan, "Robust application of new deep learning tools: an experimental study in medical imaging," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13289–13317, 2022.

[45] C. Tonkin, "Robodebt was an ai ethics disaster," 2021, https://ia.acs.org.au/article/2021/robodebt-was-an-ai-ethics-disaster.html.

[46] The Conversation, "A robot breaks the finger of a 7-year-old," 2022, https://the.conversation.com/.a-robot-breaks-the-finger-of-a-7-/year-old-a-lesson-in-the-need-for-/stronger-regulation-of-artificial-.intelligence-187612.

[47] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.

[48] H. Khosravi, S. B. Shum, G. Chen et al., "Explainable artificial intelligence in education," *Computers in Education: Artificial Intelligence*, vol. 3, Article ID 100074, 2022.

[49] A. Rawal, J. McCoy, D. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: status, challenges and perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, 2021.

[50] A. Albahri, Z. Al-qaysi, L. Alzubaidi et al., "A systematic review of using deep learning technology in the steady-state visually evoked potential-based brain-computer interface applications: current trends and future trust methodology," *International Journal of Telemedicine and Applications*, vol. 2023, Article ID 7741735, 24 pages, 2023.

[51] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Evaluating stability of post-hoc explanations for business process predictions," in *Proceedings of the Service-Oriented Computing-19th International Conference, ICSOC 2021*, vol. 13121, pp. 49–64, Dubai, United Arab Emirates, November, 2021.

[52] A. Selbst and J. Powles, ""Meaningful information" and the right to explanation," in *Proceedings of the Conference on Fairness, Accountability and Transparency*, p. 48, PMLR, Atlanta GA USA, January, 2018.

[53] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Evaluating fidelity of explainable methods for predictive process analytics," in *Proceedings of the Intelligent Information Systems- CAISE Forum 2021*, vol. 424, pp. 64–72, Melbourne, Australia, June, 2021.

[54] S. Sreedharan, S. Srivastava, and S. Kambhampati, "Using state abstractions to compute personalized contrastive explanations for ai agent behavior," *Artificial Intelligence*, vol. 301, Article ID 103570, 2021.

[55] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, Article ID 102551, 2021.

[56] B. Wickramanayake, C. Ouyang, C. Moreira, and Y. Xu, "Generating purpose-driven explanations: the case of process predictive model inspection," in *Proceedings of the Intelligent Information Systems-CAISE Forum 2022*, pp. 120–129, Leuven, Belgium, June, 2022.

[57] W. X. Lim, Z. Chen, and A. Ahmed, "The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review," *Medical and Biological Engineering and Computing*, vol. 60, pp. 1–10, 2022.

[58] Y. Huang, D. Chen, W. Zhao, Y. Lv, and S. Wang, "Deep patch learning algorithms with high interpretability for regression problems," *International Journal of Intelligent Systems*, vol. 37, no. 11, pp. 8239–8276, 2022.

[59] C. Yang, A. Rangarajan, and S. Ranka, "Global model interpretation via recursive partitioning," in *Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 1563–1570, IEEE, Exeter, UK, June, 2018.

[60] C. Moreira, Y. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, and P. Bruza, "LINDA-BN: an interpretable probabilistic approach for demystifying black-box predictive models," *Decision Support Systems*, vol. 150, Article ID 113561, 2021.

[61] D. Lyu, F. Yang, H. Kwon, W. Dong, L. Yilmaz, and B. Liu, "Tdm: trustworthy decision-making via interpretability enhancement," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 450–461, 2022.

[62] C. Reed, "How should we regulate artificial intelligence?" *Philosophical Transactions of the Royal Society A:*

Mathematical, Physical & Engineering Sciences, vol. 376, no. 2128, Article ID 20170360, 2018.

[63] B. Wickramanayake, Z. He, C. Ouyang, C. Moreira, Y. Xu, and R. Sindhgatta, "Building interpretable models for business process prediction using shared and specialised attention mechanisms," *Knowledge-Based Systems*, vol. 248, Article ID 108773, 2022.

[64] R. Sindhgatta, C. Ouyang, and C. Moreira, "Exploring interpretability for predictive process analytics," in *Proceedings of the Service-Oriented Computing- 18th International Conference, ICSOC 2020*, vol. 12571, pp. 439–447, Dubai, United Arab Emirates, December, 2020.

[65] U. Bhatt, A. Xiang, S. Sharma et al., "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, Barcelona, Spain, January, 2020.

[66] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel explainable machine learning approach for eeg-based brain-computer interface systems," *Neural Computing and Applications*, vol. 34, no. 14, pp. 11347–11360, 2022.

[67] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable deep learning: a field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–397, 2022.

[68] A. de Waal and J. W. Joubert, "Explainable bayesian networks applied to transport vulnerability," *Expert Systems with Applications*, vol. 209, Article ID 118348, 2022.

[69] C. Mao, R. Lin, D. Towey, W. Wang, J. Chen, and Q. He, "Trustworthiness prediction of cloud services based on selective neural network ensemble learning," *Expert Systems with Applications*, vol. 168, Article ID 114390, 2021.

[70] R. Srinivasan and B. San Miguel González, "The role of empathy for artificial intelligence accountability," *Journal of Responsible Technology*, vol. 9, Article ID 100021, 2022.

[71] A. Choudhury and O. Asan, "Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: exploring the perceptions of healthcare practitioners in the us," *Human Factors in Healthcare*, vol. 2, Article ID 100021, 2022.

[72] S. Sharma, Y. S. Rawal, S. Pal, and R. Dani, "Fairness, accountability, sustainability, transparency (fast) of artificial intelligence in terms of hospitality industry," in *ICT Analysis and Applications*, pp. 495–504, Springer, Berlin, Germany, 2022.

[73] M. Wieringa, "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18, Barcelona, Spain, January, 2020.

[74] B. S. Cruz and M. de Oliveira Dias, "Crashed boeing 737-max: fatalities or malpractice," *GSJ*, vol. 8, pp. 2615–2624, 2020.

[75] S. Poier, "Clean and green–the volkswagen emissions scandal: failure of corporate governance?" *Problemy Ekorozwoju*, vol. 15, no. 2, pp. 33–39, 2020.

[76] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST Special Publication 1270, Gaithersburg, MD, USA, 2022.

[77] C. M. Gevaert, M. Carman, B. Rosman, Y. Georgiadou, and R. Soden, "Fairness and accountability of ai in disaster risk management: opportunities and challenges," *Patterns*, vol. 2, no. 11, Article ID 100363, 2021.

[78] F. Königstorfer and S. Thalmann, "Ai documentation: a path to accountability," *Journal of Responsible Technology*, vol. 11, Article ID 100043, 2022.

[79] I. Rahwan, M. Cebrian, N. Obradovich et al., "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.

[80] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "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*, vol. 26, no. 4, pp. 2141–2168, 2020.

[81] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[82] D. Omeiza, H. Web, M. Jirotka, and L. Kunze, "Towards accountability: providing intelligible explanations in autonomous driving," in *Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 231–237, IEEE, Nagoya, Japan, July, 2021.

[83] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals. and it's biased against blacks, Fed," *Probation*, vol. 80, p. 38, 2016.

[84] A. Kadambi, "Achieving fairness in medical devices," *Science*, vol. 372, no. 6537, pp. 30-31, 2021.

[85] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[86] M. Madaio, L. Egede, H. Subramonyam, J. Wortman Vaughan, and H. Wallach, "Assessing the fairness of ai systems: ai practitioners' processes, challenges, and needs for support," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. 1, pp. 1–26, 2022.

[87] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[88] M. von Zahn, S. Feuerriegel, and N. Kuehl, "The cost of fairness in ai: evidence from e-commerce," *Business & information systems engineering*, vol. 64, no. 3, pp. 335–348, 2022.

[89] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez et al., "Data augmentation for fairness-aware machine learning: preventing algorithmic bias in law enforcement systems," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2302–2314, Seoul, Republic of Korea, June, 2022.

[90] M. Bogen and A. Rieke, "Help wanted: an examination of hiring algorithms, equity, and bias," *Upturn, December*, vol. 7, 2018.

[91] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 134–148, PMLR, Atlanta, GA, USA, January, 2018.

[92] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, New York, NY, USA, March, 2012.

[93] L. Oneto and S. Chiappa, "Fairness in machine learning," in *Recent Trends in Learning from Data*, pp. 155–196, Springer, Berlin, Germany, 2020.

[94] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[95] V. Grari, S. Lamprier, and M. Detyniecki, "Adversarial learning for counterfactual fairness," *Machine Learning*, vol. 112, no. 3, pp. 741–763, 2022.

[96] F. P. Santos, F. C. Santos, A. Paiva, and J. M. Pacheco, "Evolutionary dynamics of group fairness," *Journal of Theoretical Biology*, vol. 378, pp. 96–102, 2015.

[97] M. M. Khalili, X. Zhang, and M. Abroshan, "Fair sequential selection using supervised learning models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28144–28155, 2021.

[98] Y. Zheng, S. Wang, and J. Zhao, "Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models," *Transportation Research Part C: Emerging Technologies*, vol. 132, Article ID 103410, 2021.

[99] P. Besse, E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser, "A survey of bias in machine learning through the prism of statistical parity," *The American Statistician*, vol. 76, no. 2, pp. 188–198, 2022.

[100] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI: challenges and opportunities," *Business and Information Systems Engineering*, vol. 62, no. 4, pp. 379–384, 2020.

[101] Y. Chen, E. Huerta, J. Duarte et al., "A fair and ai-ready Higgs boson decay dataset," *Scientific Data*, vol. 9, pp. 31–10, 2022.

[102] European Commission, *High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*, European Commission, Brussels, Belgium, 2019.

[103] IEEE, "IEEE standard computer dictionary: a compilation of IEEE standard computer glossaries," *IEEE Std*, vol. 610, pp. 1–217, 1991.

[104] M. Shafique, M. Naseer, T. Theocharides et al., "Robust machine learning systems: challenges, current trends, perspectives, and the road ahead," *IEEE Design and Test*, vol. 37, no. 2, pp. 30–57, 2020.

[105] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 2018.

[106] J. Zhang and C. Li, "Adversarial examples: opportunities and challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2578–2593, 2020.

[107] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, https://arxiv.org/abs/1412.6572.

[108] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, Salt Lake City, UT, USA, June, 2018.

[109] S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: a survey," 2020, https://arxiv.org/abs/2007.00753.

[110] N. Akhtar, M. Jalwana, M. Bennamoun, and A. S. Mian, "Attack to fool and explain deep networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5980–5995, 2022.

[111] N. Carlini, A. Athalye, N. Papernot et al., "On evaluating adversarial robustness," 2019, https://arxiv.org/abs/1902.06705.

[112] B. dos Santos Silva, C.-T. Lee, R. Williams, B.-Y. Kuo, C.-M. Chang, and S. Muppidi, "Inline detection and prevention of adversarial attacks," Springer, Berlin, Germany, US Patent App. 16/952,494, 2022.

[113] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada, April, 2018.

[114] M. Nicolae, M. Sinn, T. N. Minh et al., "Adversarial robustness toolbox v0.2.2," 2018, https://arxiv.org/abs/1807.01069.

[115] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "Robustness in deep learning for computer vision: mind the gap?" 2021, https://arxiv.org/pdf/2112.00639.pdf.

[116] Z. Luo, C. Zhu, L. Fang, G. Kou, R. Hou, and X. Wang, "An effective and practical gradient inversion attack," *International Journal of Intelligent Systems*, vol. 37, no. 11, pp. 9373–9389, 2022.

[117] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*, OpenReview.net, New Orleans, LA, USA, May, 2019.

[118] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes?" in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019*, pp. 1045–1054, IEEE, Seoul, Korea (South), October, 2019.

[119] V. Mitra, H. Franco, R. M. Stern et al., "Robust features in deep-learning-based speech recognition," in *New Era for Robust Speech Recognition, Exploiting Deep Learning*, pp. 187–217, Springer, Berlin, Germany, 2017.

[120] F. Cartella, O. Anunciação, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: application to fraud detection and imbalanced data," in *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021) co-located with the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, vol. 2808, New York, NY, USA, February, 2021.

[121] F. Karim, S. Majumdar, and H. Darabi, "Adversarial attacks on time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3309–3320, 2021.

[122] F. Taymouri, M. L. Rosa, S. Erfani, Z. D. Bozorgi, and I. Verenich, "Predictive business process monitoring via generative adversarial nets: the case of next event prediction," in *International Conference on Business Process Management*, pp. 237–256, Springer, Berlin, Germany, 2020.

[123] D. Gursoy, O. H. Chi, L. Lu, and R. Nunkoo, "Consumers acceptance of artificially intelligent (ai) device use in service delivery," *International Journal of Information Management*, vol. 49, pp. 157–169, 2019.

[124] H. Choung, P. David, and A. Ross, "Trust in ai and its role in the acceptance of ai technologies," *International Journal of Human-Computer Interaction*, vol. 39, no. 9, pp. 1727–1739, 2022.

[125] C. Nicodeme, "Build confidence and acceptance of ai-based decision support systems-explainable and liable ai," in *Proceedings of the 2020 13th International Conference on Human System Interaction (HSI)*, pp. 20–23, IEEE, Tokyo, Japan, June, 2020.

[126] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in ai," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 10–12, 2020.

[127] A. L. Ostrom, D. Fotheringham, and M. J. Bitner, "Customer acceptance of ai in service encounters: understanding antecedents and consequences," in *Handbook of Service Science*, vol. 2, pp. 77–103, Springer, Berlin, Germany, 2019.

[128] I. Mezgár and J. Váncza, "From ethics to standards–a path via responsible ai to cyber-physical production systems," *Annual Reviews in Control*, vol. 53, pp. 391–404, 2022.

[129] S. Borau, T. Otterbring, S. Laporte, and S. Fosso Wamba, "The most human bot: female gendering increases humanness perceptions of bots and acceptance of ai," *Psychology and Marketing*, vol. 38, no. 7, pp. 1052–1068, 2021.

[130] V. Braithwaite, "Beyond the bubble that is robodebt: how governments that lose integrity threaten democracy," *Australian Journal of Social Issues*, vol. 55, no. 3, pp. 242–259, 2020.

[131] C. Campione, "The dark nudge era: cambridge analytica, digital manipulation in politics, and the fragmentation of society," Bachelor's Degree Thesis, Springer, Berlin, Germany, 2018.

[132] D. Perino, K. Katevas, A. Lutu, E. Marin, and N. Kourtellis, "Privacy-preserving ai for future networks," *Communications of the ACM*, vol. 65, no. 4, pp. 52-53, 2022.

[133] H. Berghel, "Equifax and the latest round of identity theft roulette," *Computer*, vol. 50, no. 12, pp. 72–76, 2017.

[134] C. Greene and J. Stavins, "Did the target data breach change consumer assessments of payment card security?" *Journal of Payments Strategy and Systems*, vol. 11, pp. 121–133, 2017.

[135] G. Giordano, F. Palomba, and F. Ferrucci, "On the use of artificial intelligence to deal with privacy in iot systems: a systematic literature review," *Journal of Systems and Software*, vol. 193, Article ID 111475, 2022.

[136] D. Su, H. T. Huynh, Z. Chen, Y. Lu, and W. Lu, "Re-identification attack to privacy-preserving data analysis with noisy sample-mean," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1045–1053, California, CA, USA, June, 2020.

[137] T. Lee, I. M. Molloy, and D. Su, "Protecting cognitive systems from model stealing attacks," Google Patent, New York, NY, USA, US Patent 11,023,593, 2021.

[138] H. Chen, S. U. Hussain, F. Boemer et al., "Developing privacy-preserving ai systems: the lessons learned," in *Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–4, IEEE, California, CA, USA, July, 2020.

[139] M. Rosenquist, *Defense in Depth Strategy Optimizes Security*, Intel Corporation, Santa Clara, CA, USA, 2008.

[140] G. Kaissis, A. Ziller, J. Passerat-Palmbach et al., "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473–484, 2021.

[141] T. Li, S. Xie, Z. Zeng, M. Dong, and A. Liu, "Atps: an ai based trust-aware and privacy-preserving system for vehicle managements in sustainable vanets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19837–19851, 2022.

[142] N. Chmait, D. L. Dowe, Y.-F. Li, and D. G. Green, "An information-theoretic predictive model for the accuracy of ai agents adapted from psychometrics," in *Proceedings of the International Conference on Artificial General Intelligence*, pp. 225–236, Springer, San Francisco, CA, USA, October, 2017.

[143] X. Ye, L. Zhao, and L. Wang, "Diagnostic accuracy of endoscopic ultrasound with artificial intelligence for gastrointestinal stromal tumors: a meta-analysis," *Journal of Digestive Diseases*, vol. 23, no. 5-6, pp. 253–261, 2022.

[144] C. Lin, T. Chau, C.-S. Lin et al., "Point-of-care artificial intelligence-enabled ecg for dyskalemia: a retrospective cohort analysis for accuracy and outcome prediction," *NPJ digital medicine*, vol. 5, pp. 8–12, 2022.

[145] B. Haibe-Kains, G. A. Adam, A. Hosny et al., "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, 2020.

[146] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: still a ways to go," *Science Translational Medicine*, vol. 13, no. 586, Article ID eabb1655, 2021.

[147] O. E. Gundersen and S. Kjensmo, "State of the art: reproducibility in artificial intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Washington DC, USA, February, 2018.

[148] O. E. Gundersen, S. Shamsaliei, and R. J. Isdahl, "Do machine learning platforms provide out-of-the-box reproducibility?" *Future Generation Computer Systems*, vol. 126, pp. 34–47, 2022.

[149] J. Wang, J. Jiang, D. Zhang et al., "An integrated ai model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules," *European Radiology*, vol. 32, no. 3, pp. 2120–2129, 2022.

[150] R. Koulu, "Proceduralizing control and discretion: human oversight in artificial intelligence policy," *Maastricht Journal of European and Comparative Law*, vol. 27, no. 6, pp. 720–735, 2020.

[151] I. Garcia-Magarino, R. Muttukrishnan, and J. Lloret, "Human-centric ai for trustworthy iot systems with explainable multilayer perceptrons," *IEEE Access*, vol. 7, pp. 125562–125574, 2019.

[152] R. Fanni, V. E. Steinkogler, G. Zampedri, and J. Pierson, "Enhancing human agency through redress in artificial intelligence systems," *AI & Society*, vol. 38, no. 2, pp. 537–547, 2022.

[153] B. C. Stahl, R. Rodrigues, N. Santiago, and K. Macnish, "A european agency for artificial intelligence: protecting fundamental rights and ethical values," *Computer Law and Security Report*, vol. 45, Article ID 105661, 2022.

[154] W. Liang, G. A. Tadesse, D. Ho et al., "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[155] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, "Traceability for trustworthy ai: a review of models and tools," *Big Data and Cognitive Computing*, vol. 5, no. 2, p. 20, 2021.

[156] T. Harrison, L. F. Luna-Reyes, T. Pardo, N. De Paula, M. Najafabadi, and J. Palmer, "The data firehose and ai in government: why data management is a key to value and ethics," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, pp. 171–176, New York, NY, USA, December 2019.

[157] L. Alzubaidi, J. Zhang, A. J. Humaidi et al., "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 53–74, 2021.

[158] P. Anagnostou, M. Capocasa, N. Milia et al., "When data sharing gets close to 100%: what human paleogenetics can teach the open science movement," *PLoS One*, vol. 10, no. 3, Article ID e0121409, 2015.

[159] D. Pandove and A. Malhi, "A correlation based recommendation system for large data sets," *Journal of Grid Computing*, vol. 19, no. 4, pp. 42–23, 2021.

[160] S. Chai, W. Chu, Z. Zhang, Z. Li, and M. Z. Abedin, "Dynamic nonlinear connectedness between the green bonds, clean energy, and stock price: the impact of the COVID-19 pandemic," *Annals of Operations Research*, vol. 28, 2022.

[161] L. Alzubaidi, J. Bai, A. Al-Sabaawi et al., "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, 2023.

[162] Z. Alammar, L. Alzubaidi, J. Zhang, Y. Li, W. Lafta, and Y. Gu, "Deep transfer learning with enhanced feature fusion for detection of abnormalities in x-ray images," *Cancers*, vol. 15, p. 4007, 2023.

[163] A. H. Al-Timemy, L. Alzubaidi, Z. M. Mosa et al., "A deep feature fusion of improved suspected keratoconus detection with deep learning," *Diagnostics*, vol. 13, no. 10, p. 1689, 2023.

[164] R. I. Hasan, S. M. Yusuf, M. S. Mohd Rahim, and L. Alzubaidi, "Automatic clustering and classification of coffee leaf diseases based on an extended kernel density estimation approach," *Plants*, vol. 12, no. 8, p. 1603, 2023.

[165] M. A. Shyaa, Z. Zainol, R. Abdullah, M. Anbar, L. Alzubaidi, and J. Santamaría, "Enhanced intrusion detection with data stream classification and concept drift guided by the incremental learning genetic programming combiner," *Sensors*, vol. 23, no. 7, p. 3736, 2023.

[166] F. H. Awad, M. M. Hamad, and L. Alzubaidi, "Robust classification and detection of big medical data using advanced parallel k-means clustering, yolov4, and logistic regression," *Life*, vol. 13, no. 3, p. 691, 2023.

[167] S. A. Jebur, K. A. Hussein, H. K. Hoomod, L. Alzubaidi, and J. Santamaría, "Review on deep learning approaches for anomaly event detection in video surveillance," *Electronics*, vol. 12, no. 1, p. 29, 2022.

[168] G. Abbas, A. Mehmood, M. Carsten, G. Epiphaniou, and J. Lloret, "Safety, security and privacy in machine learning based internet of things," *Journal of Sensor and Actuator Networks*, vol. 11, no. 3, p. 38, 2022.

[169] A. Strzelecki and M. Rizun, "Consumers' change in trust and security after a personal data breach in online shopping," *Sustainability*, vol. 14, no. 10, p. 5866, 2022.

[170] C. Thapa and S. Camtepe, "Precision health data: requirements, challenges and existing techniques for data security and privacy," *Computers in Biology and Medicine*, vol. 129, Article ID 104130, 2021.

[171] W. Liang, Y. Yang, C. Yang et al., "Pdpchain: a consortium blockchain-based privacy protection scheme for personal data," *IEEE Transactions on Reliability*, vol. 72, no. 2, pp. 586–598, 2023.

[172] B. van Giffen, D. Herhausen, and T. Fahse, "Overcoming the pitfalls and perils of algorithms: a classification of machine learning biases and mitigation methods," *Journal of Business Research*, vol. 144, pp. 93–106, 2022.

[173] M.-P. Fernando, F. Cèsar, N. David, and H.-O. José, "Missing the missing values: the ugly duckling of fairness in machine learning," *International Journal of Intelligent Systems*, vol. 36, no. 7, pp. 3217–3258, 2021.

[174] T. F. Kusumasari and R. Fauzi, "Design guidelines and process of metadata management based on data management body of knowledge," in *Proceedings of the 2021 7th International Conference on Information Management (ICIM)*, pp. 87–91, IEEE, London, UK, March 2021.

[175] L. R. Kaplan, M. Farooque, D. Sarewitz, and D. Tomblin, "Designing participatory technology assessments: a reflexive method for advancing the public role in science policy decision-making," *Technological Forecasting and Social Change*, vol. 171, Article ID 120974, 2021.

[176] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "Data governance taxonomy: cloud versus non-cloud," *Sustainability*, vol. 10, no. 1, p. 95, 2018.

[177] N. Thompson, R. Ravindran, and S. Nicosia, "Government data does not mean data governance: lessons learned from a public sector application audit," *Government Information Quarterly*, vol. 32, no. 3, pp. 316–322, 2015.

[178] T. Usova and R. Laws, "Teaching a one-credit course on data literacy and data visualisation," *Journal of Information Literacy*, vol. 15, no. 1, p. 84, 2021.

[179] L. Edwards and M. Veale, "Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"," *IEEE Security & Privacy*, vol. 16, no. 3, pp. 46–54, 2018.

[180] L. Ungerer and S. Slade, "Ethical considerations of artificial intelligence in learning analytics in distance education contexts," in *Learning Analytics in Open and Distributed Learning*, pp. 105–120, Springer, Berlin, Germany, 2022.

[181] C. D. Kloos, Y. Dimitriadis, D. Hernández-Leo et al., "H2o learn-hybrid and human-oriented learning: trustworthy and human-centered learning analytics (tahcla) for hybrid education," in *Proceedings of the 2022 IEEE Global Engineering Education Conference (EDUCON)*, pp. 94–101, IEEE, Tunis, Tunisia, March 2022.

[182] L. Wilton, S. Ip, M. Sharma, and F. Fan, "Where is the ai? ai literacy for educators," in *International Conference on Artificial Intelligence in Education*, pp. 180–188, Springer, Berlin, Germany, 2022.

[183] V. A. Gensini, C. Converse, W. S. Ashley, and M. Taszarek, "Machine learning classification of significant tornadoes and hail in the United States using era5 proximity soundings," *Weather and Forecasting*, vol. 36, pp. 2143–2160, 2021.

[184] C. Calvo-Sancho, J. Díaz-Fernández, Y. Martín et al., "Supercell convective environments in Spain based on era5: hail and non-hail differences," *Weather and Climate Dynamics*, vol. 3, no. 3, pp. 1021–1036, 2022.

[185] A. J. Hill and R. S. Schumacher, "Forecasting excessive rainfall with random forests and a deterministic convection-allowing model," *Weather and Forecasting*, vol. 36, pp. 1693–1711, 2021.

[186] A. McGovern, I. Ebert-Uphoff, D. J. Gagne, and A. Bostrom, "Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science," *Environmental Data Science*, vol. 1, p. e6, 2022.

[187] S. Kantayya, *Coded Bias*, 7th empire media, London, UK, 2002.

[188] S. E. Brammer, "Documentary review: coded bias," *Feminist Pedagogy*, vol. 2, p. 12, 2022.

[189] V. Mithal, G. Nayak, A. Khandelwal, V. Kumar, R. Nemani, and N. C. Oza, "Mapping burned areas in tropical forests using a novel machine learning framework," *Remote Sensing*, vol. 10, no. 2, p. 69, 2018.

[190] M. Molinaro and G. Orzes, "From forest to finished products: the contribution of industry 4.0 technologies to the

wood sector," *Computers in Industry*, vol. 138, Article ID 103637, 2022.

[191] A. Khandelwal, A. Karpatne, P. Ravirathinam et al., "Realsat, a global dataset of reservoir and lake surface area variations," *Scientific Data*, vol. 9, pp. 356–412, 2022.

[192] I. Duporge, O. Isupova, S. Reece, D. W. Macdonald, and T. Wang, "Using very-high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes," *Remote Sensing in Ecology and Conservation*, vol. 7, no. 3, pp. 369–381, 2021.

[193] D. Tuia, B. Kellenberger, S. Beery et al., "Perspectives in machine learning for wildlife conservation," *Nature Communications*, vol. 13, pp. 792–815, 2022.

[194] C. Chilson, K. Avery, A. McGovern, E. Bridge, D. Sheldon, and J. Kelly, "Automated detection of bird roosts using nexrad radar data and convolutional neural networks," *Remote Sensing in Ecology and Conservation*, vol. 5, no. 1, pp. 20–32, 2019.

[195] W. Ruan, K. Wu, Q. Chen, and C. Zhang, "Resnet-based bioacoustics presence detection technology of hainan gibbon calls," *Applied Acoustics*, vol. 198, Article ID 108939, 2022.

[196] R. M. Rogers, J. Buler, T. Clancy, and H. Campbell, "Repurposing open-source data from weather radars to reduce the costs of aerial waterbird surveys," *Ecological Solutions and Evidence*, vol. 3, Article ID e12148, 2022.

[197] D. Diochnos, S. Mahloujifar, and M. Mahmoody, "Adversarial risk and robustness: general definitions and implications for the uniform distribution," *Advances in Neural Information Processing Systems*, vol. 31, 2022.

[198] M. S. Pydi and V. Jog, "The many faces of adversarial risk," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10000–10012, 2021.

[199] G. Alicioglu and B. Sun, "A survey of visual analytics for explainable artificial intelligence methods," *Computers & Graphics*, vol. 102, pp. 502–520, 2022.

[200] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, 2021.

[201] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[202] J. A. Esterhuizen, B. R. Goldsmith, and S. Linic, "Interpretable machine learning for knowledge generation in heterogeneous catalysis," *Nature catalysis*, vol. 5, no. 3, pp. 175–184, 2022.

[203] Q. Wang, L. T. Tan, R. Q. Hu, and Y. Qian, "Hierarchical energy-efficient mobile-edge computing in iot networks," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11626–11639, 2020.

[204] H. Hu, Q. Wang, R. Q. Hu, and H. Zhu, "Mobility-aware offloading and resource allocation in a mec-enabled iot network with energy harvesting," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17541–17556, 2021.

[205] S. Fu, F. Zhou, and R. Q. Hu, "Resource allocation in a relay-aided mobile edge computing system," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23659–23669, 2022.

[206] H. Xu, P. V. Klaine, O. Onireti, B. Cao, M. Imran, and L. Zhang, "Blockchain-enabled resource management and sharing for 6g communications," *Digital Communications and Networks*, vol. 6, no. 3, pp. 261–269, 2020.

[207] J. Wang, Z. Yan, H. Wang, T. Li, and W. Pedrycz, "A survey on trust models in heterogeneous networks," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2127–2162, 2022.

[208] S. Taimoor, L. Ferdouse, and W. Ejaz, "Holistic resource management in uav-assisted wireless networks: an optimization perspective," *Journal of Network and Computer Applications*, vol. 205, Article ID 103439, 2022.

[209] H. Xu, L. Zhang, O. Onireti, Y. Fang, W. J. Buchanan, and M. A. Imran, "Beeptrace: blockchain-enabled privacy-preserving contact tracing for covid-19 pandemic and beyond," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3915–3929, 2021.

[210] R. Kumar, A. A. Khan, J. Kumar et al., "Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 16301–16314, 2021.

[211] L. Ricci, D. D. F. Maesa, A. Favenza, and E. Ferro, "Blockchains for covid-19 contact tracing and vaccine support: a systematic review," *IEEE Access*, vol. 9, pp. 37936–37950, 2021.

[212] Q. Zhang, J. Wu, M. Zanella, W. Yang, A. K. Bashir, and W. Fornaciari, "Sema-iiovt: emergent semantic-based trustworthy information-centric fog system and testbed for intelligent internet of vehicles," *IEEE Consumer Electronics Magazine*, vol. 12, no. 1, pp. 70–79, 2023.

[213] M. Abdel-Basset, N. Moustafa, H. Hawash, and W. Ding, "Federated learning for privacy-preserving internet of things," in *Deep Learning Techniques for IoT Security and Privacy*, pp. 215–228, Springer, Berlin, Germany, 2022.

[214] A. Makkar and J. H. Park, "Securecps: cognitive inspired framework for detection of cyber attacks in cyber–physical systems," *Information Processing & Management*, vol. 59, no. 3, Article ID 102914, 2022.

[215] A. Makkar, U. Ghosh, D. B. Rawat, and J. H. Abawajy, "Fedlearnsp: preserving privacy and security using federated learning and edge computing," *IEEE Consumer Electronics Magazine*, vol. 11, no. 2, pp. 21–27, 2022.

[216] S. Tarikere, I. Donner, and D. Woods, "Diagnosing a healthcare cybersecurity crisis: the impact of iomt advancements and 5g," *Business Horizons*, vol. 64, no. 6, pp. 799–807, 2021.

[217] B. Ghimire and D. B. Rawat, "Secure, privacy preserving and verifiable federating learning using blockchain for internet of vehicles," *IEEE Consumer Electronics Magazine*, vol. 11, no. 6, pp. 67–74, 2022.

[218] L. Malina, G. Srivastava, P. Dzurenda, J. Hajny, and S. Ricci, "A privacy-enhancing framework for internet of things services," in *International Conference on Network and System Security*, pp. 77–97, Springer, Berlin, Germany, 2019.

[219] M. Baza, R. Amer, A. Rasheed, G. Srivastava, M. Mahmoud, and W. Alasmary, "A blockchain-based energy trading scheme for electric vehicles," in *Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–7, IEEE, Las Vegas, NV, USA, January 2021.

[220] R. Jabbar, E. Dhib, A. B. Said et al., "Blockchain technology for intelligent transportation systems: a systematic literature review," *IEEE Access*, vol. 10, pp. 20995–21031, 2022.

[221] S. Khan, F. Luo, Z. Zhang et al., "A privacy-preserving and transparent identity management scheme for vehicular social networking," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 11555–11570, 2022.

[222] K. N. Qureshi, L. Shahzad, A. Abdelmaboud et al., "A blockchain-based efficient, secure and anonymous conditional privacy-preserving and authentication scheme for the

internet of vehicles," *Applied Sciences*, vol. 12, no. 1, p. 476, 2022.

[223] Y. Guo, Z. Wan, H. Cui, X. Cheng, and F. Dressler, "Vehicloak: a blockchain-enabled privacy-preserving payment scheme for location-based vehicular services," *IEEE Transactions on Mobile Computing*, vol. 8, pp. 1–13, 2022.

[224] W. Ahmed, W. Di, and D. Mukathe, "Privacy-preserving blockchain-based authentication and trust management in vanets," *IET Networks*, vol. 11, no. 3-4, pp. 89–111, 2022.

[225] L. T. Tan, R. Q. Hu, and L. Hanzo, "Twin-timescale artificial intelligence aided mobility-aware edge caching and computing in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3086–3099, 2019.

[226] J. Liu, M. Ahmed, M. A. Mirza et al., "Rl/drl meets vehicular task offloading using edge and vehicular cloudlet: a survey," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8315–8338, 2022.

[227] T. Alladi, V. Kohli, V. Chamola, and F. R. Yu, "A deep learning based misbehavior classification scheme for intrusion detection in cooperative intelligent transportation systems," *Digital Communications and Networks*, vol. 13, 2022.

[228] G. Muhammad and M. Alhussein, "Security, trust, and privacy for the internet of vehicles: a deep learning approach," *IEEE Consumer Electronics Magazine*, vol. 11, no. 6, pp. 49–55, 2022.

[229] T. Le and S. Shetty, "Artificial intelligence-aided privacy preserving trustworthy computation and communication in 5g-based iot networks," *Ad Hoc Networks*, vol. 126, Article ID 102752, 2022.

[230] G. Kumar, M. Lydia, and Y. Levron, "Security challenges in 5g and iot networks: a review," *Secure Communication for 5G and IoT Networks*, vol. 9, pp. 1–13, 2022.

[231] D. N. Molokomme, A. J. Onumanyi, and A. M. Abu-Mahfouz, "Edge intelligence in smart grids: a survey on architectures, offloading models, cyber security measures, and challenges," *Journal of Sensor and Actuator Networks*, vol. 11, no. 3, p. 47, 2022.

[232] K. N. Qureshi, A. Alhudhaif, M. A. Qureshi, and G. Jeon, "Nature-inspired solution for coronavirus disease detection and its impact on existing healthcare systems," *Computers and Electrical Engineering*, vol. 95, Article ID 107411, 2021.

[233] M. H. Bohara, K. Patel, A. Saiyed, and A. Ganatra, "Adversarial artificial intelligence assistance for secure 5g-enabled iot," in *Blockchain for 5G-Enabled IoT*, pp. 323–350, Springer, Berlin, Germany, 2021.

[234] J. Jenefa and E. Mary Anita, "Secure authentication schemes for vehicular adhoc networks: a survey," *Wireless Personal Communications*, vol. 123, no. 1, pp. 31–68, 2022.

[235] A. Didouh, H. Labiod, Y. E. Hillali, and A. Rivenq, "Blockchain-based collaborative certificate revocation systems using clustering," *IEEE Access*, vol. 10, pp. 51487–51500, 2022.

[236] Z. Lu, G. Qu, and Z. Liu, "A survey on recent advances in vehicular network security, trust, and privacy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 760–776, 2019.

[237] A. P. Mdee, M. M. Saad, M. Khan, M. T. R. Khan, and D. Kim, "Impacts of location-privacy preserving schemes on vehicular applications," *Vehicular Communications*, vol. 36, Article ID 100499, 2022.

[238] X. He, X. Niu, Y. Wang, L. Xiong, Z. Jiang, and C. Gong, "A hierarchical blockchain-assisted conditional privacy-preserving authentication scheme for vehicular ad hoc networks," *Sensors*, vol. 22, no. 6, p. 2299, 2022.

[239] S. Babu and A. Raj Kumar P, "A comprehensive survey on simulators, emulators, and testbeds for vanets," *International Journal of Communication Systems*, vol. 35, no. 8, Article ID e5123, 2022.

[240] M. Elaryh Makki Dafalla, R. A. Mokhtar, R. A. Saeed, H. Alhumyani, S. Abdel-Khalek, and M. Khayyat, "An optimized link state routing protocol for real-time application over vehicular ad-hoc network," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4541–4556, 2022.

[241] A. Nauman, T. N. Nguyen, Y. A. Qadri, Z. Nain, K. Cengiz, and S. W. Kim, "Artificial intelligence in beyond 5g and 6g reliable communications," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 73–78, 2022.

[242] N. M. Elfatih, M. K. Hasan, Z. Kamal et al., "Internet of vehicle's resource management in 5g networks using ai technologies: current status and trends," *IET Communications*, vol. 16, no. 5, pp. 400–420, 2022.

[243] Building Design Construction, "Robotics," 2022, https://www.bdcnetwork.com/robotics-new-way-demolish-buildings.

[244] A. Darko, A. P. Chan, M. A. Adabre, D. J. Edwards, M. R. Hosseini, and E. E. Ameyaw, "Artificial intelligence in the aec industry: scientometric analysis and visualization of research activities," *Automation in Construction*, vol. 112, Article ID 103081, 2020.

[245] Y. Pan and L. Zhang, "Roles of artificial intelligence in construction engineering and management: a critical review and future trends," *Automation in Construction*, vol. 122, Article ID 103517, 2021.

[246] B. C. Stahl and D. Wright, "Ethics and privacy in ai and big data: implementing responsible research and innovation," *IEEE Security & Privacy*, vol. 16, no. 3, pp. 26–33, 2018.

[247] M. Beltrami, G. Orzes, J. Sarkis, and M. Sartor, "Industry 4.0 and sustainability: towards conceptualization and theory," *Journal of Cleaner Production*, vol. 312, Article ID 127733, 2021.

[248] B. Wan, C. Xu, R. P. Mahapatra, and P. Selvaraj, "Understanding the cyber-physical system in international stadiums for security in the network from cyber-attacks and adversaries using ai," *Wireless Personal Communications*, vol. 127, no. 2, pp. 1207–1224, 2021.

[249] F. Yuan, E. Klavon, Z. Liu, R. P. Lopez, and X. Zhao, "A systematic review of robotic rehabilitation for cognitive training," *Frontiers in Robotics and AI*, vol. 8, Article ID 605715, 2021.

[250] H. He, J. Gray, A. Cangelosi, Q. Meng, T. McGinnity, and J. Mehnen, "The challenges and opportunities of artificial intelligence for trustworthy robots and autonomous systems," in *Proceedings of the 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE)*, pp. 68–74, IEEE, Oxford, UK, April 2020.

[251] R. E. Stuck, B. E. Holthausen, and B. N. Walker, "The role of risk in human-robot trust," in *Trust in Human-Robot Interaction*, pp. 179–194, Elsevier, Amsterdam, Netherlands, 2021.

[252] P. McAleenan, C. McAleenan, G. Ayers, M. Behm, and Z. Beachem, "The ethics deficit in occupational safety and health monitoring technologies," *Proceedings of the Institution of Civil Engineers-Management, Procurement and Law*, vol. 172, no. 3, pp. 93–100, 2019.

[253] Y. Liu, M. Habibnezhad, and H. Jebelli, "Brainwave-driven human-robot collaboration in construction," *Automation in Construction*, vol. 124, Article ID 103556, 2021.

[254] J. Garcia, G. Villavicencio, F. Altimiras et al., "Machine learning techniques applied to construction: a hybrid bibliometric analysis of advances and future directions," *Automation in Construction*, vol. 142, Article ID 104532, 2022.

[255] P. Adami, P. B. Rodrigues, P. J. Woods et al., "Effectiveness of vr-based training on improving construction workers' knowledge, skills, and safety behavior in robotic teleoperation," *Advanced Engineering Informatics*, vol. 50, Article ID 101431, 2021.

[256] T. Slaton, C. Hernandez, and R. Akhavian, "Construction activity recognition with convolutional recurrent networks," *Automation in Construction*, vol. 113, Article ID 103138, 2020.

[257] Z. Salih Ageed, S. R M Zeebaree, M. Mohammed Sadeeq et al., "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 29–38, 2021.

[258] C. Hongsong, Z. Yongpeng, C. Yongrui, and B. Bhargava, "Security threats and defensive approaches in machine learning system under big data environment," *Wireless Personal Communications*, vol. 117, no. 4, pp. 3505–3525, 2021.

[259] J. Xing and Z. Zhang, "Hierarchical network security measurement and optimal proactive defense in cloud computing environments," *Security and Communication Networks*, vol. 2022, Article ID 6783223, 11 pages, 2022.

[260] Ž. Turk, B. García de Soto, B. R. Mantha, A. Maciel, and A. Georgescu, "A systemic framework for addressing cybersecurity in construction," *Automation in Construction*, vol. 133, Article ID 103988, 2022.

[261] J. M. Wing, "Trustworthy ai," *Communications of the ACM*, vol. 64, no. 10, pp. 64–71, 2021.

[262] M. A. Ağca, S. Faye, and D. Khadraoui, "A survey on trusted distributed artificial intelligence," *IEEE Access*, vol. 10, pp. 55308–55337, 2022.

[263] G. Marcus, "The next decade in ai: four steps towards robust artificial intelligence," 2002, https://arxiv.org/abs/2002.06177.

[264] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8588–8601, 2020.

[265] K. Leino, Z. Wang, and M. Fredrikson, "Globally-robust neural networks," in *International Conference on Machine Learning*, pp. 6212–6222, Proceedings of Machine Learning Research, New York, NY, USA, 2021.

[266] Y. Chen, S. Wang, Y. Qin, X. Liao, S. Jana, and D. Wagner, "Learning security classifiers with verified global robustness properties," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 477–494, Los Angeles, CA, USA, November 2021.

[267] R. Hamon, H. Junklewitz, and I. Sanchez, *Robustness and Explainability of Artificial Intelligence*, Publications Office of the European Union, Brussels, Belgium, 2021.

[268] M. Wortsman, G. Ilharco, J. W. Kim et al., "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, New Orleans, LA, USA, December 2022.

[269] M. Casadio, E. Komendantskaya, M. L. Daggitt et al., "Neural network robustness as a verification property: a principled case study," in *International Conference on Computer Aided Verification*, pp. 219–231, Springer, Berlin, Germany, 2022.

[270] S. F. Alhashmi, S. A. Salloum, and S. Abdallah, "Critical success factors for implementing artificial intelligence (ai) projects in dubai government United Arab Emirates (uae) health sector: applying the extended technology acceptance model (tam)," in *International Conference on Advanced Intelligent Systems and Informatics*, pp. 393–405, Springer, Berlin, Germany, 2019.

[271] K. Govindan, "How artificial intelligence drives sustainable frugal innovation: a multitheoretical perspective," *IEEE Transactions on Engineering Management*, vol. 87, pp. 1–18, 2022.

[272] K. E. Henry, R. Kornfield, A. Sridharan et al., "Human-–machine teaming is key to ai adoption: clinicians' experiences with a deployed machine learning system," *NPJ digital medicine*, vol. 5, pp. 97–106, 2022.

[273] T. H. Chang, L. T. Watson, J. Larson et al., "Algorithm 1028: vtmop: solver for blackbox multiobjective optimization problems," *ACM Transactions on Mathematical Software*, vol. 48, no. 3, pp. 1–34, 2022.

[274] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, pp. 1–12, Glasgow, UK, May 2019.

[275] O. Vereschak, G. Bailly, and B. Caramiaux, "How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. 2, pp. 1–39, 2021.

[276] G. Kaptchuk, D. G. Goldstein, E. Hargittai, J. M. Hofman, and E. M. Redmiles, "How good is good enough? quantifying the impact of benefits, accuracy, and privacy on willingness to adopt covid-19 decision aids," *Digital Threats: Research and Practice*, vol. 3, no. 3, pp. 1–18, 2022.

[277] S. Cai, Z. Ma, M. J. Skibniewski, and S. Bao, "Construction automation and robotics for high-rise buildings over the past decades: a comprehensive review," *Advanced Engineering Informatics*, vol. 42, Article ID 100989, 2019.

[278] A. Al Rashid, S. A. Khan, S. G Al-Ghamdi, and M. Koc, "Additive manufacturing: technology, applications, markets, and opportunities for the built environment," *Automation in Construction*, vol. 118, Article ID 103268, 2020.

[279] S. K. Yevu, A. T. W. Yu, A. Darko, and M. N. Addy, "Evaluation model for influences of driving forces for electronic procurement systems application in ghanaian construction projects," *Journal of Construction Engineering and Management*, vol. 147, no. 8, Article ID 04021076, 2021.

[280] Q. K. Jahanger, J. Louis, D. Trejo, and C. Pestana, "Potential influencing factors related to digitalization of construction-phase information management by project owners," *Journal of Management in Engineering*, vol. 37, no. 3, Article ID 04021010, 2021.

[281] G. Ma, M. Wu, Z. Wu, and W. Yang, "Single-shot multibox detector-and building information modeling-based quality inspection model for construction projects," *Journal of Building Engineering*, vol. 38, Article ID 102216, 2021.

[282] M. Karaz, J. C. Teixeira, and K. M. Rahla, "Construction and demolition waste—a shift toward lean construction and building information model," in *Sustainability and Automation in Smart Constructions*, pp. 51–58, Springer, Berlin, Germany, 2021.

[283] S. O. Abioye, L. O. Oyedele, L. Akanbi et al., "Artificial intelligence in the construction industry: a review of present status, opportunities and future challenges," *Journal of Building Engineering*, vol. 44, Article ID 103299, 2021.

[284] M. Regona, T. Yigitcanlar, B. Xia, and R. Y. M. Li, "Opportunities and adoption challenges of ai in the construction

industry: a prisma review," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 1, p. 45, 2022.

[285] X. Xu and H. Yang, "Vision measurement of tunnel structures with robust modelling and deep learning algorithms," *Sensors*, vol. 20, no. 17, p. 4945, 2020.

[286] L. M. Dang, H. Wang, Y. Li et al., "Automatic tunnel lining crack evaluation and measurement using deep learning," *Tunnelling and Underground Space Technology*, vol. 124, Article ID 104472, 2022.

[287] N. Hoch and S. Brad, "Managing ai technologies in earthwork construction: a triz-based innovation approach," in *International TRIZ Future Conference*, pp. 3–14, Springer, Berlin, Germany, 2020.

[288] G. Marcus and E. Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, Vintage, New York, NY, USA, 2019.

[289] T. Fountaine, B. McCarthy, and T. Saleh, "Building the ai-powered organization," *Harvard Business Review*, vol. 97, pp. 62–73, 2019.

[290] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, pp. 3–25, Elsevier, Amsterdam, Netherlands, 2021.

[291] J. E. Plaks, L. Bustos Rodriguez, and R. Ayad, "Identifying psychological features of robots that encourage and discourage trust," *Computers in Human Behavior*, vol. 134, Article ID 107301, 2022.

[292] R. Stower, N. Calvo-Barajas, G. Castellano, and A. Kappas, "A meta-analysis on children's trust in social robots," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1979–2001, 2021.

[293] D. K. Singh, M. Kumar, E. Fosch-Villaronga, D. Singh, and J. Shukla, "Ethical considerations from child-robot interactions in under-resourced communities," *International Journal of Social Robotics*, vol. 882, pp. 1–17, 2022.

[294] J. Borenstein, A. R. Wagner, and A. Howard, "Overtrust of pediatric health-care robots: a preliminary survey of parent perspectives," *IEEE Robotics and Automation Magazine*, vol. 25, no. 1, pp. 46–54, 2018.

[295] C. A. Miller, "Trust, transparency, explanation, and planning: why we need a lifecycle perspective on human-automation interaction," in *Trust in Human-Robot Interaction*, pp. 233–257, Elsevier, Amsterdam, Netherlands, 2021.

[296] C. Lutz and A. Tamò-Larrieux, "Do privacy concerns about social robots affect use intentions? evidence from an experimental vignette study," *Frontiers in Robotics and AI*, vol. 8, Article ID 627958, 2021.

[297] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *Journal of human-robot interaction*, vol. 3, no. 2, p. 74, 2014.

[298] K. Dörfler, G. Dielemans, L. Lachmayer et al., "Additive manufacturing using mobile robots: opportunities and challenges for building construction," *Cement and Concrete Research*, vol. 158, Article ID 106772, 2022.

[299] D. Huang, Q. Chen, J. Huang, S. Kong, and Z. Li, "Customer-robot interactions: understanding customer experience with service robots," *International Journal of Hospitality Management*, vol. 99, Article ID 103078, 2021.

[300] E. K. Chiou and J. D. Lee, "Trusting automation: designing for responsivity and resilience," *Human Factors*, vol. 18, Article ID 00187208211009995, 2021.

[301] A. Martinho, N. Herber, M. Kroesen, and C. Chorus, "Ethical issues in focus by the autonomous vehicles industry," *Transport Reviews*, vol. 41, no. 5, pp. 556–577, 2021.

[302] B. Shneiderman, "Human-centered artificial intelligence: reliable, safe & trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[303] B. H. Guo, Y. Zou, Y. Fang, Y. M. Goh, and P. X. Zou, "Computer vision technologies for safety science and management in construction: a critical review and future research directions," *Safety Science*, vol. 135, Article ID 105130, 2021.

[304] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset," *Automation in Construction*, vol. 106, Article ID 102894, 2019.

[305] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: real-time detection of personal protective equipment," *Automation in Construction*, vol. 112, Article ID 103085, 2020.

[306] Z. Wang, Y. Zhang, K. M. Mosalam, Y. Gao, and S.-L. Huang, "Deep semantic segmentation for visual understanding on construction sites," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 2, pp. 145–162, 2022.

[307] C. Brosque, E. Galbally, O. Khatib, and M. Fischer, "Human-robot collaboration in construction: opportunities and challenges," in *Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–8, IEEE, Ankara, Turkey, June 2020.

[308] M. Wu, J.-R. Lin, and X.-H. Zhang, "How human-robot collaboration impacts construction productivity: an agent-based multi-fidelity modeling approach," *Advanced Engineering Informatics*, vol. 52, Article ID 101589, 2022.

[309] D. Nozaki, K. Okamoto, T. Mochida et al., "Ai management system to prevent accidents in construction zones using 4k cameras based on 5g network," in *Proceedings of the 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 462–466, IEEE, Chiang Rai, Thailand, November 2018.

[310] H. Baker, M. R. Hallowell, and A. J.-P. Tixier, "Ai-based prediction of independent construction safety outcomes from universal attributes," *Automation in Construction*, vol. 118, Article ID 103146, 2020.

[311] J. C. Augusto and C. D. Nugent, *Designing Smart Homes: The Role of Artificial Intelligence*, vol. 4008, Springer, Berlin, Germany, 2006.

[312] V. K. Shukla and B. Singh, "Conceptual framework of smart device for smart home management based on rfid and iot," in *Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 787–791, IEEE, Dubai, United Arab Emirates, February 2019.

[313] C.-J. Liang, X. Wang, V. R. Kamat, and C. C. Menassa, "Human–robot collaboration in construction: classification and research trends," *Journal of Construction Engineering and Management*, vol. 147, no. 10, Article ID 03121006, 2021.

[314] C. Zheyuan, M. A. Rahman, H. Tao, Y. Liu, D. Pengxuan, and Z. M. Yaseen, "Need for developing a security robot-based risk management for emerging practices in the workplace using the advanced human-robot collaboration model," *Work*, vol. 68, no. 3, pp. 825–834, 2021.

[315] D. Hornig, *Optimized safety layouts for fenceless robots*, Technische Universität Braunschweig, Braunschweig, Germany, Ph.D. Thesis, 2022.

[316] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers, and A. Sandygulova, "Perceived safety in physical human–robot interaction—a survey," *Robotics and Autonomous Systems*, vol. 151, Article ID 104047, 2022.

[317] T. P. Huck, N. Münch, L. Hornung, C. Ledermann, and C. Wurll, "Risk assessment tools for industrial human-robot collaboration: novel approaches and practical needs," *Safety Science*, vol. 141, Article ID 105288, 2021.

[318] S. You, J.-H. Kim, S. Lee, V. Kamat, and L. P. Robert, "Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments," *Automation in Construction*, vol. 96, pp. 161–170, 2018.

[319] J. M. Davila Delgado, L. Oyedele, A. Ajayi et al., "Robotics and automated systems in construction: understanding industry-specific challenges for adoption," *Journal of Building Engineering*, vol. 26, Article ID 100868, 2019.

[320] T. Kopp, M. Baumgartner, and S. Kinkel, "Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework," *The International Journal of Advanced Manufacturing Technology*, vol. 112, no. 3-4, pp. 685–704, 2021.

[321] R.-J. Halme, M. Lanz, J. Kämäräinen, R. Pieters, J. Latokartano, and A. Hietanen, "Review of vision-based safety systems for human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 111–116, 2018.

[322] S. R. Schepp, J. Thumm, S. B. Liu, and M. Althoff, "Sara: a tool for safe human-robot coexistence and collaboration through reachability analysis," in *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, pp. 4312–4317, IEEE, Philadelphia, PA, USA, May 2022.

[323] Y. Xu and H. Bao, "Fintech regulation: evolutionary game model, numerical simulation, and recommendations," *Expert Systems with Applications*, vol. 211, Article ID 118327, 2023.

[324] X. Vives, *The Impact of Fintech on Banking*, European Economy, Brussels, Belgium, 2017.

[325] S. Biswas, B. Carson, V. Chung, S. Singh, and R. Thomas, *Ai-bank of the Future: Can banks Meet the Ai challenge*, McKinsey & Company, New York, NY, USA, 2020.

[326] A. Hanif, "Towards explainable artificial intelligence in banking and financial services," 2022, https://arxiv.org/abs/2112.08441.

[327] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: an application to default risk analysis," 2019, https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

[328] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser et al., "Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[329] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, 2022.

[330] S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy, responsible ai," *Frontiers in Artificial Intelligence*, vol. 5, Article ID 779799, 2022.

[331] C. Maree, J. E. Modal, and C. W. Omlin, "Towards responsible ai for financial transactions," in *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 16–21, IEEE, Canberra, Australia, December 2020.

[332] S. D. Rosadi, S. Yuniarti, and R. Fauzi, "Protection of data privacy in the era of artificial intelligence in the financial sector in Indonesia," *Journal of Central Banking Law and Institutions*, vol. 1, pp. 353–366, 2022.

[333] J.-H. Chen, Y.-J. Wang, Y.-C. Tsai, and S. Y.-C. Chen, "Financial vision based differential privacy applications," 2021, https://arxiv.org/abs/2112.14075.

[334] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[335] H. Surendra and H. Mohan, "A review of synthetic data generation methods for privacy preserving data publishing," *International Journal of Scientific & Technology Research*, vol. 6, pp. 95–101, 2017.

[336] R. Max, A. Kriebitz, and C. Von Websky, "Ethical considerations about the implications of artificial intelligence in finance," *Handbook on Ethics in Finance*, vol. 18, pp. 577–592, 2021.

[337] H. Allahabadi, J. Amann, I. Balot et al., "Assessing trustworthy ai in times of covid-19. deep learning for predicting a multi-regional score conveying the degree of lung compromise in covid-19 patients," *IEEE Transactions on Technology and Society*, vol. 3, no. 4, pp. 272–289, 2022.

[338] G. Karimian, E. Petelos, and S. M. Evers, "The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review," *AI and Ethics*, vol. 2, no. 4, pp. 539–551, 2022.

[339] C. González-Gonzalo, E. F. Thee, C. C. Klaver et al., "Trustworthy ai: closing the gap between development and integration of ai systems in ophthalmic practice," *Progress in Retinal and Eye Research*, vol. 90, Article ID 101034, 2022.

[340] Z. H. Khan, A. Siddique, and C. W. Lee, "Robotics utilization for healthcare digitization in global covid-19 management," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, p. 3819, 2020.

[341] E. J. MacKay and M. D. Stubna, "Understanding basic concepts of supervised machine learning model development in the clinical setting," *Journal of Cardiothoracic and Vascular Anesthesia*, vol. 35, no. 8, pp. 2336-2337, 2021.

[342] M. Braun, P. Hummel, S. Beck, and P. Dabrock, "Primer on an ethics of ai-based decision support systems in the clinic," *Journal of Medical Ethics*, vol. 47, no. 12, p. e3, 2021.

[343] American Medical Association, *Code of Medical Ethics*, American Medical Association, Philadelphia, PA, USA, 1848.

[344] J. D. Shahidullah, C. A. Hostutler, and S. G. Forman, "Ethical considerations in medication-related roles for pediatric primary care psychologists," *Clinical Practice in Pediatric Psychology*, vol. 7, no. 4, pp. 405–416, 2019.

[345] A. Sheikh, M. Anderson, S. Albala et al., "Health information technology and digital innovation for national learning health and care systems," *The Lancet Digital Health*, vol. 3, no. 6, pp. e383–e396, 2021.

[346] N. A. Smuha, "The eu approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.

[347] T. Gundersen and K. Bærøe, "The future ethics of artificial intelligence in medicine: making sense of collaborative models," *Science and Engineering Ethics*, vol. 28, no. 2, pp. 17–16, 2022.

[348] L. Vesnic-Alujevic, S. Nascimento, and A. Polvora, "Societal and ethical impacts of artificial intelligence: critical notes on

european policy frameworks," *Telecommunications Policy*, vol. 44, no. 6, Article ID 101961, 2020.

[349] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension," *BMJ: British Medical Journal*, vol. 370, 2020.

[350] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022.

[351] A. Esteva, K. Chou, S. Yeung et al., "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, pp. 5–9, 2021.

[352] K. Cusi, S. Isaacs, D. Barb et al., "American association of clinical endocrinology clinical practice guideline for the diagnosis and management of nonalcoholic fatty liver disease in primary care and endocrinology clinical settings: cosponsored by the american association for the study of liver diseases (aasld)," *Endocrine Practice*, vol. 28, no. 5, pp. 528–562, 2022.

[353] A. D. Pearson, C. Rossig, C. Mackall et al., "Paediatric strategy forum for medicinal product development of chimeric antigen receptor t-cells in children and adolescents with cancer: accelerate in collaboration with the european medicines agency with participation of the food and drug administration," *European Journal of Cancer*, vol. 160, pp. 112–133, 2022.

[354] A. Fiocchi, R. Pawankar, C. Cuello-Garcia et al., "World allergy organization-mcmaster university guidelines for allergic disease prevention (glad-p): probiotics," *World Allergy Organization Journal*, vol. 8, pp. 4–13, 2015.

[355] I. S. W. Group, "Software as a medical device": possible framework for risk categorization and corresponding considerations," in *International Medical Device Regulators Forum*, Springer, Berlin, Germany, 2014.

[356] A. Adeyemo, M. K. Balaconis, D. R. Darnes et al., "Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps," *Nature Medicine*, vol. 27, no. 11, pp. 1876–1884, 2021.

[357] P. Galetsi, K. Katsaliaki, and S. Kumar, "Exploring benefits and ethical challenges in the rise of mhealth (mobile healthcare) technology for the common good: an analysis of mobile applications for health specialists," *Technovation*, vol. 121, Article ID 102598, 2023.

[358] E. Fosch-Villaronga, H. Drukarch, P. Khanna, T. Verhoef, and B. Custers, "Accounting for diversity in ai for medicine," *Computer Law and Security Report*, vol. 47, Article ID 105735, 2022.

[359] M. DeCamp and C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 2020–2023, 2020.

[360] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[361] E. Ntoutsi, P. Fafalios, U. Gadiraju et al., "Bias in data-driven artificial intelligence systems—an introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Article ID e1356, 2020.

[362] K. Ahmad, M. Maabreh, M. Ghaly, K. Khan, J. Qadir, and A. Al-Fuqaha, "Developing future human-centered smart cities: critical analysis of smart city security, data management, and ethical challenges," *Computer Science Review*, vol. 43, Article ID 100452, 2022.

[363] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing bias in artificial intelligence in health care," *JAMA*, vol. 322, no. 24, pp. 2377-2378, 2019.

[364] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022.

[365] S. Faghani, B. Khosravi, K. Zhang et al., "Mitigating bias in radiology machine learning: 3. performance metrics," *Radiology: Artificial Intelligence*, vol. 4, no. 5, Article ID e220061, 2022.

[366] D. A. Vyas, L. G. Eisenstein, and D. S. Jones, "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms," *New England Journal of Medicine*, vol. 383, no. 9, pp. 874–882, 2020.

[367] A. Khan, M. C. Turchin, A. Patki et al., "Genome-wide polygenic score to predict chronic kidney disease across ancestries," *Nature Medicine*, vol. 28, no. 7, pp. 1412–1420, 2022.

[368] B. Mittelstadt, "Principles alone cannot guarantee ethical ai," *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[369] A. Felländer, J. Rebane, S. Larsson, M. Wiggberg, and F. Heintz, "Achieving a data-driven risk assessment methodology for ethical ai," *Digital Society*, vol. 1, no. 2, pp. 13–27, 2022.

[370] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of Biomedical Informatics*, vol. 113, Article ID 103621, 2021.

[371] L. L. Guo, S. R. Pfohl, J. Fries et al., "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine," *Scientific Reports*, vol. 12, pp. 2726–2810, 2022.

[372] G. Mudd-Martin, A. L. Cirino, V. Barcelona et al., "Considerations for cardiovascular genetic and genomic research with marginalized racial and ethnic groups and indigenous peoples: a scientific statement from the American heart association," *Circulation: Genomic and Precision Medicine*, vol. 14, no. 4, Article ID e000084, 2021.

[373] M. Livingston, "Preventing racial bias in federal ai," *JSPG*, vol. 16, no. 2, 2020.

[374] A. H. Sham, K. Aktas, D. Rizhinashvili et al., "Ethical ai in facial expression analysis: racial bias," *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 399–406, 2022.

[375] V. Baxi, R. Edwards, M. Montalto, and S. Saha, "Digital pathology and artificial intelligence in translational medicine and clinical practice," *Modern Pathology*, vol. 35, no. 1, pp. 23–32, 2022.

[376] R. Benjamin, "Race after technology: abolitionist tools for the new jim code," *Social Forces*, vol. 98, no. 4, pp. 1–3, 2020.

[377] European Group on Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics And'autonomous' Systems: Brussels*, EU: European Union, Brussels, Belgium, 2018.

[378] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[379] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[380] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–31, 2023.

[381] D. Mahapatra, A. Poellinger, and M. Reyes, "Interpretability-Guided inductive bias for deep learning based medical image," *Medical Image Analysis*, vol. 81, Article ID 102551, 2022.

[382] E. S. Ho and Z. Ding, "Electrocardiogram analysis of post-stroke elderly people using one-dimensional convolutional neural network model with gradient-weighted class activation mapping," *Artificial Intelligence in Medicine*, vol. 130, Article ID 102342, 2022.

[383] B. Jiang, Y. Zhang, L. Zhang, G. H de Bock, R. Vliegenthart, and X. Xie, "Human-recognizable ct image features of subsolid lung nodules associated with diagnosis and classification by convolutional neural networks," *European Radiology*, vol. 31, no. 10, pp. 7303–7315, 2021.

[384] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability," *International Journal of Information Management*, vol. 69, Article ID 102538, 2023.

[385] L. Alzubaidi, O. Al-Shamma, M. A. Fadhel, L. Farhan, J. Zhang, and Y. Duan, "Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model," *Electronics*, vol. 9, no. 3, p. 445, 2020.

[386] A. Deshpande and H. Sharp, "Responsible ai systems: who are the stakeholders?" in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 227–236, Oxford, UK, May 2022.

[387] K. Liu and D. Tao, "The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services," *Computers in Human Behavior*, vol. 127, Article ID 107026, 2022.

[388] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an ai device human-like? the role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior*, vol. 122, Article ID 106855, 2021.

[389] V. Venkatesh, J. Y. Thong, and X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS Quarterly*, vol. 36, no. 1, pp. 157–178, 2012.

[390] P. R. Daugherty and H. J. Wilson, *Human+ Machine: Reimagining Work in the Age of AI*, Harvard Business Press, Harvard, MA, USA, 2018.

[391] N. Haefner, J. Wincent, V. Parida, and O. Gassmann, "Artificial intelligence and innovation management: a review, framework, and research agenda," *Technological Forecasting and Social Change*, vol. 162, Article ID 120392, 2021.

[392] X. Wang and R. Zhou, "Impacts of user expectation and disconfirmation on satisfaction and behavior intention: the moderating effect of expectation levels," *International Journal of Human-Computer Interaction*, vol. 39, no. 15, pp. 3127–3140, 2022.

[393] Q. Yang, A. Scuito, J. Zimmerman, J. Forlizzi, and A. Steinfeld, "Investigating how experienced ux designers effectively work with machine learning," in *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 585–596, Hong Kong, China, June 2018.

[394] L. Alzubaidi, M. A. Fadhel, O. Al-Shamma et al., "Towards a better understanding of transfer learning for medical imaging: a case study," *Applied Sciences*, vol. 10, no. 13, p. 4523, 2020.

[395] H. Subramonyam, C. Seifert, and E. Adar, "Towards a process model for co-creating ai experiences," 2021, https://arxiv.org/abs/2104.07595.

[396] A. Rechkemmer and M. Yin, "When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–14, New Orleans, LA, USA, April 2022.

[397] B. Thuraisingham, "Trustworthy machine learning," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 21–24, 2022.

[398] D. Llorente, M. Ballesteros, I. D. J. S. Ramos, and J. I. C. Oria, "Deep learning adapted to differential neural networks used as pattern classification of electrophysiological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, 2022.

[399] S. A. Jebur, K. A. Hussein, H. K. Hoomod, and L. Alzubaidi, "Novel deep feature fusion framework for multi-scenario violence detection," *Computers*, vol. 12, no. 9, p. 175, 2023.

[400] R. M. Hazarbassanov, L. Al-Zubaidi, Z. M. Mosa et al., "The suitability of color histogram-based features for keratoconus detection from corneal thickness with and neural networks," *Investigative Ophthalmology and Visual Science*, vol. 64, p. 1089, 2023.

[401] L. Alzubaidi, Y. Duan, A. Al-Dujaili et al., "Deepening into the suitability of using pre-trained models of imagenet against a lightweight convolutional neural network in medical imaging: an experimental study," *PeerJ Computer Science*, vol. 7, p. e715, 2021.

[402] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi et al., "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, p. 1590, 2021.