Trustworthy Artificial Intelligence: A Review

DAVINDER KAUR, SULEYMAN USLU, KALEY J. RITTICHIER, and ARJAN DURRESI,
Indiana University–Purdue University Indianapolis

Artificial intelligence (AI) and algorithmic decision making are having a profound impact on our daily lives. These systems are vastly used in different high-stakes applications like healthcare, business, government, education, and justice, moving us toward a more algorithmic society. However, despite so many advantages of these systems, they sometimes directly or indirectly cause harm to the users and society. Therefore, it has become essential to make these systems safe, reliable, and trustworthy. Several requirements, such as fairness, explainability, accountability, reliability, and acceptance, have been proposed in this direction to make these systems trustworthy. This survey analyzes all of these different requirements through the lens of the literature. It provides an overview of different approaches that can help mitigate AI risks and increase trust and acceptance of the systems by utilizing the users and society. It also discusses existing strategies for validating and verifying these systems and the current standardization efforts for trustworthy AI. Finally, we present a holistic view of the recent advancements in trustworthy AI to help the interested researchers grasp the crucial facets of the topic efficiently and offer possible future research directions.

CCS Concepts: • **Computing methodologies → Cross-validation**; **Intelligent agents;**

Additional Key Words and Phrases: Artificial intelligence, machine learning, black-box problem, trustworthy AI, explainable AI, fairness, explainability, accountability, privacy, acceptance

ACM Reference format:
Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* 55, 2, Article 39 (January 2022), 38 pages.
https://doi.org/10.1145/3491209

1 INTRODUCTION

Artificial intelligence (AI) and algorithmic decision making are transforming our lives. In today's world, most of our day-to-day tasks are either done or guided by machines or algorithms. This area of algorithmic decision making is not new. We have been using machines for decision making for a long time. However, today these systems have become very efficient and complex because of the availability of vast data, advanced algorithms, and high computing power. It has

This work was partially supported by the National Science Foundation (NSF) under grant 1547411 and by the U.S. Department of Agriculture (USDA), National Institute of Food and Agriculture (NIFA) (award 2017-67003-26057) via an interagency partnership between USDA-NIFA and the NSF on the research program Innovations at the Nexus of Food, Energy, and Water Systems.
Authors' address: D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, Indiana University-Purdue University Indianapolis, Computer & Information Science, 723 W Michigan St, Indianapolis, IN 46202; emails: {davikaur, suslu, krittich}@iu.edu, adurresi@iupui.edu.

0360-0300/2022/01-ART39 $15.00
https://doi.org/10.1145/3491209

become challenging to interpret the logic of these systems, which sometimes makes it difficult to assess these systems properly. Because of these systems' limitations, biases, and ethical issues, they can become brittle and unfair.These AI systems hugely rely on the data to make decisions. Sometimes the data on which these systems are trained has errors and biases, making them unfair. For example, the CalGang database, a crime dataset used to predict violent crime related to a gang, is found to be extremely skewed and riddled with errors leading to bias and unfairness [46]. The recidivism algorithm used in U.S. courts to predict the probability of re-offending was biased against black people [4]. Amazon's recruitment algorithm was found to be sexist [50]. Besides the training data, these AI systems also rely on the continuous flow of data, making the privacy and governance of data critical to protect it from any malicious activities. A famous example is the Equifax data breach that exposed the personal data of millions of users [15]. The advanced algorithms behind these AI systems have made them complex and uninterpretable, making it hard to understand the reasoning behind the decisions. This has created a sense of distrust and less acceptance of these systems. For example, the adoption of AI-based medical diagnosis support systems among healthcare professionals is relatively low because of its uninterpretability even though it can be beneficial in day-to-day clinical practices [62]. Another concern related to these systems is the responsibility and accountability of the harm caused by them. To answer all of these concerns and to prevent the harm caused by these systems, several organizations have proposed different methods and guidelines to make AI safe, reliable, and trustworthy.

Lately, the field of trustworthy AI has been gaining attention from the government and different scientific communities. The International Organization for Standardization (ISO), an organization that works on technical, industrial, and commercial standardization, has presented different approaches to establish trust in AI systems using the properties of fairness, transparency, accountability, and controllability [92]. The European Union (EU) proposed ethical guidelines for trustworthy AI to govern and facilitate the development and working of AI systems [60]. The EU also passed a law called General Data Protection Regulation (GDPR), which gives individuals the "right to explanations" for AI decisions [206]. The National Institute of Standards and Technology (NIST) proposed a framework to measure and increase user trust in AI systems [148]. The U.S. Government Accountability Office (GAO) published a framework for the accountability and responsible use of AI [149]. The Defense Advanced Research Project Agency (DARPA) [81] also launched a program known as Explainable Artificial Intelligence (XAI), whose motive was to make these AI systems explainable and trustworthy. The involvement of all these significant institutes to make AI trustworthy shows how vital trustworthiness is for both the success of AI systems and the safety of users and society. Gartner estimates that 30% of all AI-based digital products will require the use of a trustworthy AI framework by 2025 [31], and 86% of users will trust and remain loyal to companies that use ethical AI principles [11]. These examples demonstrate the current necessity to develop AI systems using a trustworthy framework.

With the growing need for trustworthy AI, many different methods and frameworks have been proposed recently. Various methods focus on different stages of the AI lifecycle to make AI systems reliable and trustworthy. Some approaches focus on the design phase of the AI systems, which helps lay out the trustworthy requirements and expectations for AI systems. Some methods deal with the data collection, protection, and pre-processing phase, making data fair, diverse, and secure. Some approaches focus on the modeling phase of the AI system to provide explainability and interpretability of the system. Other methods work with the implementation and oversight phase of the AI system, which utilizes proper auditing and testing techniques to ensure accountability and reliability. The EU [60] stated the importance of human involvement in making AI trustworthy. Some researchers also proposed the notion of collaborative intelligence, which uses both humans and machines for decision making [213]. The common objective of all of these methods is to ensure

that AI systems behave as intended without causing harm to the users or society, thus leading to trust in the systems.

Motivated by the recent concerns with AI systems and the need for trustworthy AI, this article presents an in-depth review of trustworthy AI requirements and associated methods. We offer a holistic view of the recent advancements in trustworthy AI to help interested researchers grasp the crucial facets of the topic efficiently and present possible directions for future research. In this article, we have tried to answer the following research questions:

- *R1:* What are the requirements to make AI trustworthy?
- *R2:* What guidelines and policies are required to govern the working of AI systems?
- *R3:* Why is human involvement significant in this changing era of AI?
- *R4:* What aspects are essential to make AI decisions acceptable?

In this article, we make three contributions. First, we present a comprehensive background, concepts, and need for a trustworthy AI system. Second, we review and organize the existing methods and guidelines that make AI systems trustworthy. We have labeled different proposed methods for trustworthy AI requirements with the level of human involvement needed and where they can be implemented and reviewed in the AI lifecycle. Last we compare different proposed methods based on the trustworthy requirements they fulfill. In this article, we also discuss additional verification and validation techniques to test trustworthy AI systems.

The article is summarized as follows. Section 2 presents an overview of the foundational concepts related to traditional and trustworthy AI. Section 3 presents a survey of the available methods and organizes them based on trustworthy requirements. Section 4 focuses on the available verification and validation techniques to test trustworthy AI systems. Section 5 highlights the current challenges and future directions. Finally, Section 6 concludes the article.

2 BACKGROUND AND FOUNDATIONAL CONCEPTS

This section presents essential background and foundational concepts. Section 2.1 offers essential definitions related to trustworthy AI, Section 2.2 describes problems with the traditional AI and the need for trustworthy AI, Section 2.3 presents requirements of trustworthy AI, and Section 2.4 focuses on the human-centered approach of trustworthy AI.

2.1 Preliminaries and Definitions

Trustworthy AI is not a monolithic concept but a polylithic one [100]. Different terms in this field have several different interpretations. Therefore, it is imperative to define and explain these terms before we can use them. This section contains essential definitions of the terms related to the field of AI:

- `Artificial Intelligence`: AI is a field that deals with making machines think. Legg and Hutter [122] define AI as a process of imitating human behavior and decision-making capabilities. So, AI is a way to train machines to perform tasks that require intelligence.
- `Black-Box Problem`: The black-box problem means that the system is opaque, and it is difficult to track the structure, internal working, and system implementation [2]. AI systems are becoming more complicated, making them challenging to understand [37]. This problem decreases the system's trustworthiness as it is challenging to provide the reasoning and explanation for the output.
- `Explainable and Interpretable AI`: Explainable and interpretable AI deals with developing explainable and interpretable models. Miller [139] defines explainable AI as how an explanatory agent provides reasoning for their own or another agent's decision

making. Arrieta et al. [5] describe explainable AI as a suite of algorithmic techniques that generate high-performance explainable models that humans can easily understand and trust. Researchers often use the terms *explainability* and *interpretability* interchangeably [114]. So, in this article, we also use them interchangeably.

- `Reliability`: Reliability of the system ensures that the system performs as intended—that is, within specified limits and without any failure, it produces the same outputs for the same inputs consistently [92].
- `Fairness`: Fairness of the system ensures that there is an absence of any discrimination or favoritism toward an individual or a group [136] based on any inherent or acquired characteristics that are irrelevant in the context of decision making [177].
- `Accountability`: Accountability refers to the need to explain and justify the actions and decisions made by the system to different users with whom the system interacts [211].
- `Privacy`: Privacy makes sure that the sensitive data that is either shared by an individual or collected by an AI system is protected from any unjustifiable or illegal gathering and use of data [92].
- `Acceptance`: Acceptance of an AI system is the user's willingness to use the system in service encounters [82].
- `Trust`: Trust is a complex phenomenon [53]. Different disciplines defined trust differently. Sociologists view trust as an attribute of human relationships [79] and psychologists consider it as a cognitive attribute [167], whereas economists think it is calculative [212]. An agreement among these definitions is that trust has something to do with integrity and reliability. Philosophically, the National Institute of Standards and Technology [23] defines trust as "the confidence one element has in another, that second element will behave as expected."
- `Trustworthy AI`: Trustworthy AI is a framework to ensure that a system is worthy of being trusted based on the evidence concerning its stated requirements. It makes sure that the users' and stakeholders' expectations are met in a verifiable way [92].

2.2 Need for Trustworthy AI

AI systems are rapidly transforming every aspect of life from movie recommendations to diagnosing diseases, assisting customers, and much more [213]. With enormous applications of AI, this rapid development has also raised many concerns. The late Stephen Hawking once said that "AI impact can be cataclysmic unless its rapid development is controlled" [193]. AI systems can be dangerous and harmful if strict measures are not followed in designing and overseeing them [131]. In today's world, numerous sectors are utilizing AI systems in decision making, but these AI systems do not always yield good results. Their usefulness comes with a great responsibility of making sure that they do not cause any harm to humanity. However, sometimes these AI systems failed and showed dangerous consequences for humans. The COMPAS algorithm used across the nation to predict the risk of criminal recidivism is found to be biased against black people [4]. A facial recognition software tagged black people with inappropriate labels because of the low quality of sample data used to train the system [141]. Resume screening used by a major tech company was biased against women [50]. These examples show how bias can mislead the black-box system and cause harm or unfairness. These systems have sometimes even caused harm by behaving unreliably. For example, a self-driving car killed a pedestrian on the road when its algorithm malfunctioned and did not respond when its sensors detected a pedestrian in the way [115]. Furthermore, the complexity of these systems hinders the understanding of the reasoning behind decisions, hence preventing them from being used to their full potential. For example, Fan et al. [62] showed how the adoption of AI-based medical diagnosis support systems among healthcare professionals is relatively low even though they can be very useful in day-to-day clinical practices. This was found to be due to

the uninterpretable nature of these systems, hence decreasing their trust and acceptance among doctors. All of these examples show how important it is to make AI systems safe and trustworthy.

These days, AI systems have achieved enough performance to be used widely in our society. These technologies are already transforming people's lives [61]. However, even though these AI systems have some utility, this does not imply that they are good enough and trustworthy. This informal attitude toward these systems is inappropriate when dealing with high-stakes applications where one wrong decision can lead to dangerous consequences. These systems can be brittle and unfair. Marcus and Davis [130] provide an excellent example of facial recognition software that explains the need for trustworthy AI. If the facial recognition software is used for auto-tagging people in social media pictures, less reliable software is acceptable. Still, the same tool is unacceptable if the police want to use it to find suspects in surveillance photos. This example demonstrates how people adopt AI systems only when there are no life-critical consequences for them and society. To deliver AI benefits to high-stakes applications and increase AI systems' adoption, we need an ethical framework to control and govern them. The following section discusses different requirements needed to make AI systems safe, reliable, and trustworthy.

2.3 Requirements to Make AI Trustworthy

Over the past few years, research institutes, private organizations, and government agencies have proposed various guidelines and frameworks to make AI trustworthy [60, 68, 69, 154, 198, 224]. However, the sheer volume of these proposed principles has led to confusion and difficulty in agreeing upon a common set of principles to make AI trustworthy. To avoid inconsistency, some researchers [83, 95] have reviewed and analyzed the proposed principles to assess their convergence over some set of agreed upon principles. They found there to be an emerging convergence around the five main principles: transparency/explainability, justice and fairness, non-maleficence/societal and environmental well-being, responsibility/accountability, and privacy. These principles are more frequent in the proposals than other principles. Therefore, to abide by this analysis and to follow one of the first frameworks from a government organization, we have selected the EU framework [60] of trustworthy AI that has all five of these principles and also focuses on the human aspect of AI. This framework is described in the following.

The EU [60] presented three guidelines that should be followed while designing and developing the AI systems to make them trustworthy: lawful, ethical, and robust. Lawful means that the AI system's development, deployment, and use should follow all the applicable laws and regulations. Ethics implies that the AI system should respect the ethical principles and guidelines of humans. Robust means that the AI system should be technically robust while being ethical and lawful. These guidelines lay out a general framework that should be followed while developing, deploying, and using AI systems. To abide by these guidelines and make the AI system trustworthy, four ethical principles composed of seven essential requirements were proposed by the EU [60], which are summarized in Figure 1 [102, 117]. The first principle is respect for human control, ensuring that the AI system should complement humans without replacing them. The second principle is the prevention of harm, which makes sure that the AI performs as intended and does not cause any damage to the system or society. The third principle is fairness, which ensures that all social groups are treated equally without any discrimination. Last, explicability makes sure that the AI system is transparent and interpretable. The seven requirements for these four ethical principles are explained next:

- `Human agency and oversight`: AI systems should complement and empower humans without replacing them [51]. The autonomy of AI systems should be based on the risk and impact of incorrect decisions on the users and society at large. This requirement makes sure

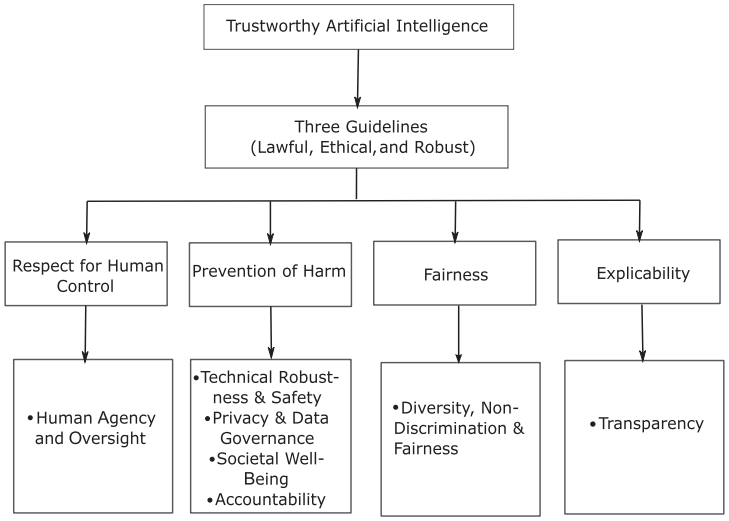

Fig. 1. Trustworthy AI framework [102, 117].

that humans are involved in the decision-making process and that this level of involvement is based on the risk and the societal and environmental impact factor [183].

- `Technical robustness and safety`: AI systems should be technically robust and perform as expected by the users [60]. This means that if something goes wrong, it should recover from the failure without causing harm to anyone. Furthermore, it should be robust enough to deal with errors at any step of the AI lifecycle. This requirement also ensures that the AI system is resilient to outside attacks and the results are reproducible.
- `Privacy and data governance`: AI systems should protect user data and govern its usage at every step of its lifecycle. This requirement makes sure that sensitive data shared by the users and collected by the AI system is protected. Furthermore, an AI system should follow all the data protection laws and regulations like GDPR [78] for the legitimate use of the data.
- `Transparency`: The transparency of an AI system refers to the need to explain, interpret, and reproduce its decisions [54]. It ensures that the different stakeholders using or impacted by the system clearly understand its performance and limitations [181].
- `Diversity, non-discrimination, and fairness`: AI systems should treat all sections of society fairly without discriminating based on factors such as socio-economic determinants. They should not cause any direct or indirect discrimination to any group of society [67]. This requirement enables the AI system to be available and accessible to all sections of the society without any discrimination.
- `Societal and environmental well-being`: AI systems should not cause any harm to society or the environment during their design, development, and use [60].
- `Accountability`: AI systems should be able to justify their decisions. This requirement deals with setting up a proper mechanism to assign responsibilities for all the correct and incorrect decisions made by the AI system [211]. It also enforces that the system is audited regularly to prevent any harm caused by it.

This list provides essential requirements for designing trustworthy AI systems. In this article, we review a total of five requirements and their implementations. Out of five requirements, four are the EU's core principles: Fairness, Explainability, Accountability, and Privacy, and based on

our research, we argue that a new principle is needed in addition: Acceptance—for incorporating a mechanism to assess AI systems based on the expectations and requirements of the users. Other trustworthy AI requirements like technical robustness, safety, and non-maleficence are beyond the scope of this work because we believe that each of these requirements has its own vast literature, which needs a separate review.

Trustworthy requirements are essential to make AI systems safe and reliable. However, some researchers [48] have argued that these requirements are inadequate, saying that they are insufficient to address the justice challenges presented by AI in society, as government and wealthy organizations can comply with these ethics requirements and still perform inequitable and unjust practices. To prevent such results, they proposed a human-focused good data approach to guide and politically approach AI development, implementation, and governance using the four pillars: Community, Rights, Useability, and Politics. Community involvement ensures the participation of the individuals and collectives in the decision-making process rather than relying on the entities in power. The rights pillar is for the AI system to abide by human rights, and its impacts should be studied. The usability pillar ensures that the access and control of the data are in the hands of the community to bring value to interpersonal relationships. Last, the politics pillar is to empower the community to lead to better activism and policy. All of these good data pillars ensure that the well-being of the people and environment should be at the forefront of trustworthy AI considerations. This proposed framework presents the importance of community and human involvement to make AI safe and trustworthy.

So, to make AI safe and reliable, both trustworthy requirements and human participation are essential. To follow this, in this article, we focus both on trustworthy requirements and human involvement to offer the readers the dimension of collaborative (Human + AI) decision making. The next section describes how humans can participate in AI decision making and the different levels of human involvement.

2.4 A Human-Centered Approach to Make AI Trustworthy (Human + AI)

The new era of AI is moving toward collaborative thinking, which is an amalgam of humans' cognitive ability and machines' exceptional computing power [51]. This new AI wave ensures that AI systems are developed to empower human beings without replacing or threatening them. Humans and machines should work as collaborative partners to achieve a goal. Humans are designing, training, deploying, and testing these systems using their cognitive abilities. Simultaneously, machines provide humans exceptional computational power, enabling the processing and analysis of data in real time [51]. A great example of collaborative human-AI work is the protein-folding AI AlphaFold, which builds on the work of hundreds of researchers. Using advanced algorithms and high computational power, AlphaFold is able to predict the structure of proteins [32]. This computational prediction takes a few days rather than the traditional several years of trial and error when done manually in the lab.

Different researchers have proposed the need to make AI systems user centric. The EU [60] has proposed human agency and oversight as a primary requirement for making AI trustworthy. Humans should be present to develop efficient algorithms, set limits for performance, flag and correct errors raised by the system, override wrong decisions, and improve the performance by giving continuous feedback. Other researchers [43] proposed a framework of human involvement in the AI lifecycle, which offers three different levels of human involvement based on the application requirements and the associated risk. The first level is for high-stakes applications like medical applications, where AI systems should only be used to assist human decisions. The second level deals with humans' medium-level involvement, which applies to applications like resume sorting and requires immediate implementation of decisions. In these types of applications, human

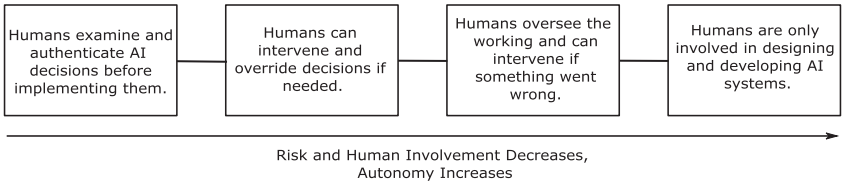

Fig. 2. Different levels of human involvement in making AI systems trustworthy.

experts should be able to view and override AI decisions. The third level deals with the low-level involvement of humans. Human experts can oversee the working of the AI system and interfere if the system starts misbehaving. Figure 2 summarizes different levels of human involvement based on the risk associated. Before discussing other human involvement methods, let us first discuss different types of users or stakeholders involved in the AI lifecycle.

Humans play a crucial role in the success of the AI system. Different users satisfy different purposes in AI systems [5]. Data scientists, researchers, and developers deal with the design, development, and continuous improvement of the system by increasing performance, correcting errors, and adding new functionalities. The second type of humans involved are the users who are directly or indirectly affected by the decisions of the AI system. These users should be able to understand the reasoning that led to a particular decision. The third type of users is the domain experts, such as doctors who use AI systems to assist in their decision making concerning their patients. These types of users should trust the AI system to be able to use it effectively. The fourth type is the policymakers and regulatory bodies, ensuring that the AI system complies with ethical and moral principles through proper auditing and testing. The last type of humans involved is managers and company board members who oversee AI system development and assess its use and application in the real world. All of these users are involved at different levels of the AI lifecycle.

To manage the involvement of different users, the EU [60] proposed human participation in three phases: human-in-the-loop, human-over-the-loop, and human-before-the-loop to make AI trustworthy:

- Human-before-the-loop deals with the methods that are applied to the design phase of the lifecycle. In these methods, humans are involved in planning, designing, and creating expectations and requirements for the AI system. In this phase, stakeholders like developers, policymakers, domain experts, and users are involved in laying out their expectations and requirements for the AI system.
- Human-in-the-loop deals with the actual development of the AI system. Humans are involved in data collection, model development, testing, and deployment of the AI system. In this phase, data scientists and developers are responsible for data collection, pre-processing, model development, and model evaluation. Policymakers and regulators can be involved through auditing mechanisms. Finally, domain experts and users are involved in the proper deployment of the AI system.
- Human-over-the-loop deals with overseeing mechanisms where humans, specifically developers, develop oversight mechanisms and re-assess the system's performance. Users or domain experts provide feedback and override the decisions when needed. Finally, policymakers make policies to govern the working of AI systems.

To evaluate the system at different levels and achieve controllability, the ISO standard [92] has proposed integrating various control points into the AI lifecycle. These control points can be used to assess the effectiveness of different trustworthy requirements at various stages of the AI lifecycle. To abide by the proposal, we have introduced four control points in the AI lifecycle:

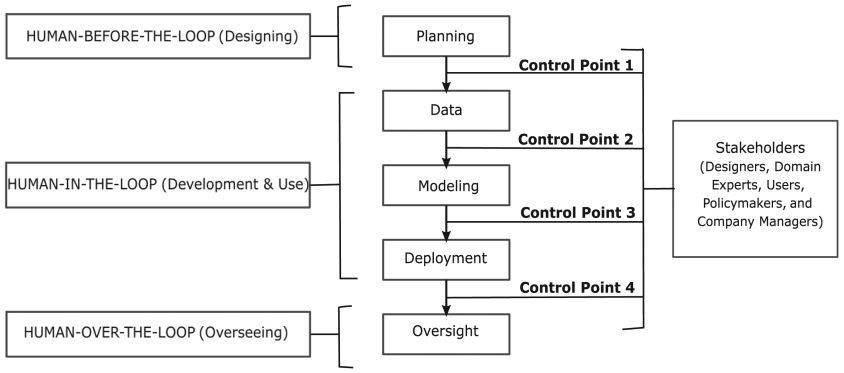

Fig. 3. Different levels of human involvement and different control points that can be used for better controllability and checking in the development of trustworthy AI.

- Control point 1 checks whether all planning requirements are met. This control point involves human-before-the-loop, where they can assess the expectations and requirements of the AI system.
- Control point 2 is used to check if proper data collection and processing were done to ensure that the data is diverse and unbiased. This control point involves human-in-the-loop, where different stakeholders can evaluate the quantity and quality of data to make it appropriate for the modeling process.
- Control point 3 is used to ensure that the correct modeling algorithm and objective function are used. Here, humans are involved in-the-loop in evaluating the model and checking if it uses appropriate attributes to make decisions.
- Control point 4 checks if proper testing and validation have been done for the AI system. Different stakeholders are involved over-the-loop at this control point to ensure that the system is adequately tested and fulfills the application requirements.

These control points are very useful in pinpointing the cause of the error in the system. To summarize, we have designed Figure 3, which presents the control points and levels of human involvement that can be used in the AI lifecycle to review the effectiveness of trustworthy AI methods. Section 3 reviews the methods for trustworthy AI requirements, categorizes them based on human involvement, and assigns different control points.

3 TRUSTWORTHY AI REQUIREMENTS

This section discusses the need, proposed methods, and technical challenges of five trustworthy AI requirements: Fairness, Explainability, Accountability, Privacy, and Acceptance.

3.1 Fairness

With the growing impact of AI systems in our daily lives, it becomes essential to ensure that these systems are fair—that is, free from any bias and discrimination. A system is called *fair* if it does not cause discrimination against any individual or group of the society [67]. However, different examples from the past have shown that these systems can be unfair if not designed, developed, implemented, and overseen correctly. Furthermore, this principle of fairness should work within ethical and moral values to ensure that these systems do not cause harm to anyone affected by them. Various studies have been conducted in this field of fairness, which are reviewed in the following sections.

3.1.1 Need of Fairness. Algorithms and AI systems use a vast amount of data and logic to perform a task and are made to ease the decision-making process. Still, they can lead to bias and unfairness if not designed, developed, and implemented correctly. Researchers analyzed these systems and found some of them to behave unfairly. Some examples include the following. A risk assessment tool used by the judicial system to predict future criminals was biased against black people [4]. The hiring algorithm used by a major tech company was found to be biased against women [20]. Some researchers [42] showed how the predictive analytic tool used to make decisions in child maltreatment screening could be biased against poor individuals or individuals of a particular race or ethnicity. Before focusing on the bias mitigation measures, there is a need to understand different causes that can lead to bias and discrimination; these causes are discussed in the following section.

3.1.2 Types of Bias. To mitigate bias and unfairness in AI systems, it is vital to understand both the biases and the factors causing them. Biases can be introduced at any step of the AI lifecycle. It can be presented by the system designers if the development team is not diverse enough, by the training data or algorithms themselves, and even by the users of the system in the implementation phase. We have summarized the different reasons for bias into three types:

- *Data bias:* The first reason for the bias to occur is if the data on which the system is trained is biased. This happens when the data does not represent a clear picture of reality. This type of bias is called *data bias*. It occurs if the training data has representation or societal bias [150]. Representation bias means that the data does not represent all segments of society equally. Societal bias happens if the data correctly represents the society that is already biased. For example, the AI system predicts that men are more suitable for engineering jobs than women because of gender stereotypes in society, reflected in the system's training data. Data bias also occurs if the population demographics of the training dataset are entirely different from the target population on which the system is being implemented [195]. Other reasons for data bias can be mismatched datasets, unlearned data cases, and manipulated datasets [165]. Therefore, to develop a fair AI system, it is crucial to make the data fair and diverse.
- *Model bias:* The second reason for the bias to occur is if the algorithm itself introduces it. Model bias can occur because of a wrong objective function that does not capture the fundamental logic for the prediction. For example, Obermeyer et al. [147] showed how the system designed to predict the need for treatment based on the level of patient illness is found to be favorably biased toward healthier white patients than sick black patients because the system was predicting the need for care based on access to healthcare benefits. This shows how the wrong choice of predictors and metrics can lead to bias in the system. Another reason for this type of bias is if one feature is given more priority than other features without any valid logic by the algorithm [21]. It is imperative to carefully select objective functions and features that do not cause any discrimination to the users or society.
- *Evaluation bias:* The third reason for bias is if the wrong evaluation metrics were used to evaluate the model. This type of bias is called *evaluation bias*. Buolamwini and Gebru [30] show how popular facial recognition systems are biased against dark-skinned females because of the unsuitable evaluation metrics used to test and validate the system. Another reason can be biased users, which affect the system through feedback loops if the users' responses are partial toward any particular item due to societal stereotypes or peer pressure [136]. Therefore, it is crucial to carefully select evaluation metrics that can detect different types of bias in the system.

3.1.3 Types of Fairness. With the increasing need for algorithmic fairness, different researchers have proposed definitions of fairness. However, there is no clear consensus on one

single definition of fairness applied to every case. The definition of fairness is context dependent, varying from application to application. Suggested definitions for fairness fall under two categories: individual fairness and group fairness. Individual fairness ensures that similar types of individuals get similar predictions [57]. Fairness definitions like fairness through awareness [57], fairness through unawareness [72], and counterfactual fairness [118] deal with individual fairness. Group fairness makes sure that all groups of the society are treated equally without any discrimination [176]. Demographic parity [118], Equalized odds [7], Equal opportunity [86], and Conditional statistical parity [44] fairness definitions deal with group fairness. A detailed discussion and comparison of fairness definitions are presented in other works [136, 205].

3.1.4 Proposed Solutions. To make AI systems fair and unbiased, several methods and techniques are offered. This section describes the proposed methods to make AI systems fair and how humans are involved in designing, developing, and overseeing these systems. We categorize the bias mitigation solutions into pre-processing, in-processing, and post-processing models. We also discuss bias detection methods and fairness toolkits.

Pre-processing models. These methods deal with pre-processing the data to make it free from any bias and discrimination. These methods ensure that the data does not over-represent or under-represent any section of the society and that it represents a clear picture of reality. Several pre-processing methods to remove bias from the data have been proposed in the literature. Brunet et al. [27] proposed a bias mitigation method for word embedding, which approximates the effect of removing a small sample of training data based on the bias of the resulting system. This system is helpful to trace the origin of the bias in the dataset. This method provides individual fairness as well as group fairness. It can work with multivariate variables. Calmon et al. [33] proposed a probabilistic framework to transform data, which will help prevent discrimination while preserving the utility of the data. Kamiran and Calders [97] discussed different data sampling and re-weighting techniques for bias prevention. These techniques reduce bias by trading off accuracy. Ruggieri [169] proposed a t-closeness technique for discrimination prevention. This technique is helpful to clean data of historical decisions before using it and provides a formal guarantee about the level of discrimination present in the dataset. Mehrabi et al. [135] proposed a method to incorporate loosely connected nodes and communities in small numbers to prevent bias and increase accuracy. This method helps to include all minority communities into the analysis, hence reducing discrimination. Luong et al. [128] proposed a technique that assigns discriminated/non-discriminated labels to the historical decisions made by the system before training the system again for bias detection and prevention. This method is helpful to detect and prevent bias from existing AI systems. Feldman et al. [63] proposed a technique of hiding protected attributes from the training dataset while still preserving the data properties. Samadi et al. [172] suggested a method that modifies input feature representations of the training data using dimensionality reduction to prevent bias. Backurs et al. [10] developed a fair clustering method that provides equal distribution of different groups in clusters to avoid bias. Sablayrolles et al. [170] proposed a technique called *radioactive data labeling*, which labels the training dataset images with an identifiable mark to ensure biased data traceability. This is a valuable method to prevent bias as it checks if the system was trained on diverse data or not. All of these methods deal with making the training data fair and diverse in the pre-processing data phase.

In-processing models. These methods prevent and mitigate bias by modifying the decision-making algorithms. Various in-processing methods have been proposed. Zafar et al. [223] proposed a method to build fair classifiers using decision boundaries, ensuring fairness concerning one or more sensitive attributes. This method maximizes fairness for given accuracy constraints and can assure accuracy to given fairness constraints. Berk et al. [16] proposed a method to make

regression algorithms fair by applying and weighting a regularizer to the standard loss function. This method provides both individual and group fairness and can calculate a numerical value for the effect of fairness on accuracy. Kamishima et al. [99] add a regularizer to the objective function to make classification algorithms fair and accurate. This method is also capable of trading-off fairness with accuracy according to user requirements. Zhang et al. [225] proposed a bias mitigation method using adversarial learning, which uses the concept of maximizing predictor accuracy while minimizing the ability to predict protected attributes. This method is helpful for both classification and regression tasks. Beaudouin et al. [13] suggest incorporating a bias mitigation process during the learning phase by introducing a penalty term to achieve a similar false-positive and false-negative rate. Kamiran et al. [98] offer bias mitigation by minimizing the information gain of protected attributes. Quadrianto and Sharmanska [156] used the concept of privileged learning—that is, using protected/sensitive attributes only in the learning phase to mitigate unfairness. This method can be applied to achieve any fairness. Huang and Vishnoi [91] designed a framework to make classification algorithms both fair and stable. They used a regularization term, which increases the accuracy while slightly decreasing the fairness. All of these methods provide fairness during the learning phase of the algorithms.

Post-processing models. These methods deal with mitigating bias by using the output of the predictors through post-processing. Various post-processing methods have been proposed using the output of the system to make it fair and unbiased. In the work of Hardt et al. [86], the method removes bias by adjusting the learned predictor to balance among supervised learning methods. This method mitigates bias while preserving the privacy of the system. Bolukbasi et al. [21] suggest modifying the unfair word embedding from the learned model to remove bias from the system. Corbett-Davies et al. [44] proposed applying a single threshold value to remove bias from all of the groups. Dwork et al. [58] proposed a decoupling technique that uses different classifiers for different groups to mitigate bias. This approach is beneficial if all of the groups are not represented equally in the data. Menon and Williamson [137] suggest mitigating bias by applying different threshold values to the objective function of various classes.

Bias detection methods. Some researchers have also presented bias detection methods to test whether the system is biased. Agarwal et al. [3] proposed a test generation mechanism to detect bias in the system. It detects all combinations of protected and non-protected attributes that can lead to discrimination through directed and undirected search. Srivastava and Rossi [187] proposed a third-party rating mechanism to detect bias using sets of biased and unbiased data. This method is capable of detecting data and algorithmic bias. Black et al. [19] proposed a fairness testing approach known as the Flip Test, which tests fairness at the individual level to detect statistical and casual discrimination.

Fairness toolkits. To ease the process of bias detection and mitigation, researchers have designed fairness toolkits. Bellamy et al. [14] developed an open source toolkit to use fairness algorithms in an industrial setting. This toolkit is designed to bring together researchers designing the algorithms and data scientists using them in the field. It contains bias detection, mitigation, and explanation algorithms that are ready to use interactively. This toolkit is very helpful, as it provides a platform to experiment and compare different bias detection and mitigation techniques. Saleiro et al. [171] created an audit toolkit to test modules for other kinds of bias present in population subgroups. In this toolkit, bias detection is done before model selection so that if the data is biased, it can be corrected before training. This toolkit will help data scientists and policymakers avoid harm by making calculated decisions about the AI models.

Different methods can be used to detect and mitigate bias. However, some researchers have raised a concern that fairness cannot be achieved until all of the stakeholders are involved in designing the system. To overcome this issue, several proposed methods incorporate different

Table 1. Fairness: Summarizes Proposed Methods for Fairness Based on the Type of Bias They Can Prevent

Level of Human Involvement	Control Points	Bias Type	Algorithm Type	Reference
Before-the-loop	1	Data/Model Bias	Any Predictor	[123, 226]
In-the-loop	2	Data Bias	Word Embedding	[27]
			Supervised Learning	[33]
			Data Pre-Processing	[63, 97, 128]
			Item-Set Mining	[169]
			Dimensionality Reduction	[172]
		Data/Model Bias	Community Detection	[135]
			Data Labeling	[170]
In-the-loop	3	Data/Model Bias	Classification	[13, 91, 99, 223]
			Regression	[16]
			Classification, Regression	[225]
			Decision Tree	[98]
Over-the-loop	4	Data/Model Bias	Classification	[86]
			Word Embedding	[21]
			Any Predictor	[3, 19, 44, 58, 137, 187, 207]

Each method is labeled based on the level of human involvement needed and control points where they can be reviewed for their effectiveness, as described in Figure 3.

stakeholders to make AI systems fair. Lepri et al. [123] proposed a multidisciplinary team of scientists, lawmakers, industry practitioners, and end users who work together in designing, developing, and evaluating AI systems to make them fair and trustworthy. They proposed the vetting of algorithms through the OPAL (Open Algorithms) project, which provides a technological and socio-political platform for a multidisciplinary team of experts to make algorithms unbiased and transparent collaboratively. Some researchers [226] argued that the definition of fairness could differ between domains involved in decision making. For instance, the definition of fairness for researchers can be different from that of end users or policymakers. To overcome this limitation, they proposed a common optimization technique for fairness and utility of the algorithm by weighting and prioritizing the different types of stakeholder fairness and model utility in the design phase itself. All of these methods aim to make the AI system fair based on the application requirements.

All of these methods to prevent and mitigate bias in the AI systems are summarized in Table 1. Different bias detection and prevention methods are organized based on the type of bias they avoid and the level of human involvement needed. We also assigned control points to all of these methods, which will help evaluate and report the effectiveness of the methods at different levels in the AI lifecycle. For example, if a fairness method is assigned control point 2, the method is applied at the data pre-processing phase and evaluated for its effectiveness later. Suppose that some stakeholders are not satisfied with the method's performance. In that case, they can make an informed decision to re-run the same method again or try another method at either the same or a different control point to reach their target fairness. This evaluation at different control points will provide developers and policymakers controllability to achieve a fair AI system.

3.1.5 Technical Challenges. Many methods and definitions of fairness have been proposed to detect and mitigate bias in AI systems. However, selecting a single definition and method to detect and mitigate all types of bias is not simple. as one definition can forego the other. Feuerriegel et al. [64] proposed that research is needed for the definition and perception of fairness, which could depend on the context of AI applications. A particular attribute can be considered sensitive for some applications and not for others. There is a need for frameworks and policies to define fairness according to the context of applications clearly. Another challenge in this field is that the definition of fairness can differ between stakeholders. There is a need for diverse multi-stakeholder

involvement in making AI fair and trustworthy. Last, more robust testing strategies and policies are needed to detect and prevent different biases in the system.

3.2 Explainability

Algorithms and AI systems make some of our day-to-day decisions. To trust these decisions, it becomes essential for different stakeholders involved with these systems to understand the reasons that led to a decision. However, these models have become overly complicated, which makes explaining them a real issue. Various studies have been conducted in this field of explainability to make these systems transparent. The following sections discuss the need for explainability, different types of explanations, methods of explanation, and evaluation metrics.

3.2.1 Need of Explainability. Explainability is used to communicate the reasoning for the AI system's decisions to different stakeholders. There is a need to explain AI systems' decisions to increase the user's trust in the system. If system users clearly understand the reasons that lead to a particular output and the cases where this system will not work, they tend to trust the decisions more [180]. The explainability of an AI system helps ensure that the decisions made by the AI system are correct. It also helps the system designers detect unknown vulnerabilities and correct errors and policymakers to design better laws to govern the system. With audiences having different levels of expertise, the explanations must be tailored to their expertise and application requirements [1]. Furthermore, the GDPR act [78] provides users the "right to explanations" for the output of the AI system. Some of the explanations that different entities seek to understand the system better are listed next:

- Explanation of how the AI system arrived at a particular decision
- Explanation of the training data on which the results are based
- Explanation of the metrics used to measure the validity or invalidity of the results.

All of these explainability questions will lead to the transparency and interpretability of opaque AI systems that are hard to interpret and understand. Explainability will help justify the predictions, improve the models, gain new insights, and lead to better governance of the AI system.

3.2.2 Types of Explanations. Different kinds of explanations can be provided based on the kind of users to whom an explanation is to be given and the application requirements. One way to distinguish explanations is based on their level of interpretability. Researchers [2] have categorized them into two types: global and local interpretability. Global interpretability methods deal with explaining the whole logic and working of AI systems. It provides a global picture of the model and reasoning for its possible outcomes. It is mainly used for applications that predict global population trends like climate change [219]. Because of the scale, global interpretability is challenging to achieve in practice. Local interpretability methods are more widely used and deal with explaining a particular decision made by the system. This interpretability is instance based and used to provide explanations of specific decisions made by the model.

Another way to differentiate explanations is based on when they are provided to the stakeholders. Considering this criterion, explanations are divided into two categories [160]: Ex-ante and Ex-post explanations. Ex-ante explanations are the ones about the use, working, and features of the AI system that are given to different stakeholders before the actual use. The purpose of Ex-ante explanations is to establish the initial trust in the system. This explanation assures that the system is well designed, tested, and validated. Ex-post explanations deal with explaining the features and circumstances that lead to a particular decision. After getting a decision from the system, these explanations are given and used to validate the initial trust established by Ex-ante explanations.

According to ISO [92], both Ex-ante and Ex-post explanations are essential for the transparency of the AI system.

All of these different explanations have a common goal: to provide interpretability and transparency to opaque AI models, increasing users' trust. The following section describes different explainability approaches that have been proposed in the literature.

3.2.3 Proposed Solutions. Different methods have been proposed in the literature to make AI systems transparent and explainable. In this section, we describe pre-modeling, in-modeling, and post-modeling explainability methods and categorized them based on their level of involvement in the AI lifecycle.

Pre-modeling approaches. One way to provide transparency and explainability to the AI system is by explaining the datasets on which the system is trained. These pre-modeling approaches deal with exploring and understanding the datasets before developing the model. They provide Ex-ante explanations to the AI system. Various studies presented approaches to make data explainable. Some researchers [89, 179, 182] proposed visualization techniques to better understand the data before using it. These techniques will help the designers and developers of the system better understand the distribution of various attributes in the dataset. Another way to achieve explainable data is through data standardization. Holland et al. [90] proposed a method of data labeling, in which data is labeled based on its quantitative and qualitative properties. This method is useful to assess the fitness of data for the application more quickly. Another similar approach [75] is creating a data sheet consisting of all information related to the data collection process, dataset features, and recommended use. These dataset standardization methods like labeling and data sheet creation will facilitate communication and understanding between different entities using them, providing transparency and explainability to the AI system.

In-modeling approaches. Another way to provide explainability is by making interpretable models. These approaches deal with giving explanations for the decisions made by the AI system. In-modeling approaches work well with less complicated model families like decision trees [153], linear models [173], and rule-based models [71]. These explainability approaches are model specific, meaning that they can only be applied to a specific family of models.

One way to provide interpretability is through tree-based ensemble models [40]. These models utilize the graph structure of trees where internal nodes represent tests on features and leaf nodes represent class labels. Different paths from the root to leaf nodes represent different interpretable classification rules. Another way to provide interpretability is through decision/rule sets [120], which use association rules like an if-then rule [120] or m-of-n rule [142] to generate classification rules. The main difference between tree-based and rule-based methods is that the tree method provides graphical interpretability, whereas the rule sets provide textual interpretability. Researchers have also proposed using linear models [87] to provide interpretability by visualizing the weight and sign of the features for a given output. This means that if the weight is high and the sign is positive, it will increase the output of the model. The disadvantage of using these interpretable models is that they are only usable when the size of the classification rules and the dimensionality of the features are within a human-understandable range.

Post-modeling approaches. Another way to reach explainability is by building proxy models on top of black-box/complex models. This type of explainability approach is applied to non-interpretable AI models. Most of the proposed methods lie in this category. We have categorized these approaches into four categories: feature importance explainability approaches, example-based explainability approaches, rule-based explainability approaches, and visualization-based explainability approaches, which are explained next.

Feature importance explainability approaches: These methods provide explainability by assigning feature importance values to the input variables. These values reflect which features played a more critical role in the decision-making process. Various feature importance explanation methods have been proposed in the literature. Ribeiro et al. [161] proposed an explanation method called *LIME* (Local Interpretable Model-agnostic Explanations), which provides local interpretability to different classifiers and regressors by highlighting essential features that led to the decision. Still, it requires the data to be converted into binary form for human interpretability. Other researchers [196] proposed a similar approach that provides explanations based on the formal requirements using the Monte Carlo algorithm. This approach helps provide explanations using simple logical rules. Some other researchers [9] proposed a technique called *LRP* (Layerwise Relevance Propagation) for image classification algorithms, which includes interpretability by computing every pixel's contribution to the prediction made by the classifier. This method utilizes heat maps to visualize the pixel importance. Other researchers [227] proposed a method for visual explanations using the information encoded in feature vectors. An automated concept-based explanation (ACE) method to provide human-understandable explanations was developed in the work of Ghorbani et al. [77]. Fisher et al. [66] suggested the MCR (Model Class Reliance) method for calculating feature importance to provide interpretability. The SHAP (SHapely Additive exPlanations) method [127] provides interpretability by assigning feature scores to each attribute for different predictions.

Example-based explainability approaches: These methods provide explanations and interpretability by creating proxy examples of the model and are based on selecting some instances from the input dataset and monitoring their corresponding outputs to explain the system. Several example-based approaches have been proposed in the literature. Kim et al. [109] proposed methods for selecting different types of instances from the data that capture all of its characteristics. To provide explainability to the system, these instances are then used as sample inputs to build prototypes. In another work, Kim et al. [108] developed a method called *MMD-critic*, which uses both prototype building and criticism selection to provide human-level interpretability. This method claims that prototypes are not enough to give interpretability to complex black boxes. As humans learn by questioning, so they proposed a criticism-based prototype model for better explainability to humans. Wachter et al. [206] proposed a method for providing counterfactual explanations that describe only the most essential variable that led to a decision and how a slight change in that variable can lead to a completely different outcome. This method is based on "what-if" scenarios and provides local explainability to the models. Mothilal et al. [140] proposed a technique that generates diverse counterfactual explanations to provide interpretability to the system.

Rule-based explainability approaches: These methods provide explainability by extracting useful information from the model and are usually applied to artificial neural networks, where useful information is extracted using the hidden layers to provide interpretability. Hailesilassie [84] reviewed various rule-extraction techniques to give comprehensibility to artificial neural networks. These techniques provide a comprehensive description of the knowledge learned in the training phase using the system's input and output. Model compression is another way of rule extraction, where complex/deep networks are compressed into simple/shallow networks to provide interpretability. Other researchers [88, 218] proposed methods for model compression to provide explainability. Kim et al. [110] proposed a method called *TCAV* (Testing with Concept Activation Vectors), which uses the internal state of neural networks to provide interpretability.

Visualization-based explainability approaches: Another method to provide explainability is visualization. These techniques offer explainability by visualizing the internal working of opaque AI systems. Some of these techniques are proposed in the literature. Casalicchio et al. [36] proposed

Table 2. Explainability: Summarizes Proposed Methods Based on the Level of Explainability Provided by Them, and Each Solution Is Labeled with the Level of Human Involvement and Control Points as Described in Figure 3

Level of Human Involvement	Control Points	Explanation Method	Scope of Explanation	Reference
Before-the-loop	1, 2	Data Visualization	Global	[179, 182]
		Data Standardization	Global	[75, 90]
In-the-loop	3	Interpretable Model	Global	[40, 87, 120]
		Feature Importance	Local	[161]
			Global, Local	[9, 66, 77, 93, 127, 227]
		Example Based	Local	[140, 206]
			Global, Local	[108, 109]
		Rule Based	Global	[88, 218]
			Global, Local	[110]
		Visualization Based	Global, Local	[36, 209, 228]
Over-the-loop	4	Evaluation Methods	Global, Local	[103, 162, 203]

a method that explains by visualizing how the change in the feature importance affects the model performance. Other researchers [228] have proposed visualizing the contribution of the evidence for the final decision to provide explainability to classifiers. However, these techniques may not be applied to very complex models because the visual representation of those models can be challenging to interpret.

All of these methods are used to make AI systems explainable. Table 2 has summarized all of these methods based on the level of human involvement needed, type of explainability provided by them, and method of explainability generation. We have also assigned control points, which will help to evaluate and report the effectiveness of these explainability methods at different levels of the development and deployment process.

3.2.4 Evaluation of Explanations. As algorithms are getting more involved in our society, different stakeholders are engaged at different AI lifecycle points. These stakeholders, such as users, domain experts, policymakers, or people affected by the decisions, require different levels of explanations. For example, domain experts designing the system require more in-depth explanations concerning which attributes are leading to a particular decision; policymakers, however, need explanations about the behavior of the system as a whole to verify if it complies with the current laws; and a user may only want to know the main reason for a particular decision. There is no single approach that can be used to satisfy the differing requirements of stakeholders. Some researchers have organized available explainability approaches based on the stakeholder's needs. Arya et al. [6] organized these approaches in a decision tree based on what is explained, the level of the explanations, and how the explanations are generated. Other researchers proposed different evaluation metrics that measure the usefulness of the explainability methods based on the stakeholder needs. ISO [92] proposed multiple measures on which explanations should be evaluated. They are consistency, continuity, and selectivity. Continuity means that similar predictions have similar explanations. Consistency means that any change in the importance of input variables is reflected in the feature score of the explanation method. Last, selectivity means that the explainable method should select the highest impact feature from the feature space successfully. Sokol and Flach [186] proposed a fact sheet for systematic evaluation of interpretable approaches. This fact sheet evaluates explainable models based on their functional, operational, usability, safety, and validation dimensions. All of these approaches help stakeholders select the best explainability approach according to their needs.

The goal of explainability is to generate trust and assist in the adoption and acceptance of AI systems. Explanations are multidimensional, which means that they cannot be attained by a single

explainability method. Additionally, given the limits to human reasoning based on their expertise to understand them [206], there is a need to generate human-centric explainability approaches based on their profiles and need. Ribera and Lapedriza [162] proposed a user-centered approach that takes into account the quality (correct information), quantity (right amount of abstraction), and relation (right relevance) according to the user's needs. Sanneman and Shah [175] proposed the concept of situation awareness based on perception, comprehension, and projection to achieve "enough explainability." They suggested that reaching "enough explainability" means that the system should include explanations for its decisions, such as why it behaved in a certain way and how model changes can affect the outcome. Other researchers [103, 203] proposed explainability acceptance metrics that utilize trust mechanisms to measure users' understanding of the explanations. All of these approaches help to evaluate explanation methods.

3.2.5 Technical Challenges. Much research has been done on generating explanations but significantly less in communicating explanations effectively and evaluating how well different users understood the explanations. Another concern is that most of the explainability approaches are used by the designers and developers of the system to debug and oversee them. There is a need to develop more explainability methods for external users who do not have any expertise to bridge the gap between transparency and actual implementation [17]. In addition, some organizations are reluctant to implement explanatory methods because of security and privacy concerns. So, appropriate methods and policies should be developed to provide the explainability of AI systems while preserving the privacy of these systems.

3.3 Accountability

There is a need to monitor the development and operation of AI decision-making algorithms to ensure that they do not cause any harm. To track them, it is vital to discover who is responsible for the harm caused by the algorithms because they are just computer programs fed with data that cannot take responsibility for their decisions. This is where accountability comes into the picture. Algorithmic accountability includes assessing the algorithms based on various parameters and assigning responsibilities of harm to different stakeholders involved in developing the algorithms. Wieringa [211] defines accountability as a networked account where responsibility is distributed among stakeholders and applied at various stages of the AI lifecycle. Several methods have been proposed to develop accountability measures for these systems; before discussing those methods, we first discuss the need for accountability.

3.3.1 Need of Accountability. Because of their use in various high-stake applications, there is a growing need for accountability measures for algorithmic decision making. It has become essential to govern these algorithmic models' design, development, and implementation to ensure they are safe and reliable. In the past, some failures in these systems have led to severe harm and life-threatening consequences. For example, a Boeing plane crashed because of some significant glitches in the computer software of the aircraft that claimed the lives of 346 people [47]. Volkswagen's new electric car software was found to have serious software architecture problems [155]. A face recognition system is biased against females and dark-skinned people [165]. These failures can be avoided with proper governance of the algorithms and holding people accountable for these failures, but the central question is who is responsible for these failures. Is it the system developers, the data collectors who provided biased data, or the domain experts using the system? The answer to all of these questions is vague until the proper mechanism for assigning responsibilities and governing the whole AI lifecycle is developed. Research on developing accountability measures is discussed in the next section.

3.3.2 Proposed Solutions. Researchers have proposed many different ways to provide accountability to an algorithmic decision-making process. The proposals include methods that can be embedded in the design process of algorithms, transparency methods, and strict laws and policies for better governance of algorithms. This section categorizes these methods into three types: Ex-ante, In-ante, and Post-ante methods.

Ex-Ante methods. These methods are specified before the actual development of the algorithms and mainly deal with the algorithms' planning and design phase, which helps assign responsibilities for the decisions made by the algorithm even before its actual development. They also deal with clearly describing all users directly or indirectly affected by the system. Several researchers proposed different methods to answer these specifications of the models effectively. McQuillan [134] proposed a prioritizing method to assign different priorities to different algorithm values through discussion between various stakeholders and the public council. This method will help algorithms resolve conflicting issues through prioritization, which prevents harm and will lead to better governance and accountability. The one drawback of this type of method is that it is difficult for people with different interests to agree on a single design. It is also hard to keep track of stakeholders' varying opinions and decisions for accountability purposes.

Other researchers [25, 29] have proposed that laying out all the design specifications and clearly describing what the system is intended to do in different circumstances enforces better governance and accountability. Broeders et al. [26] proposed a time frame mechanism in the design process that prevents harm by timely reevaluating the system to ensure it is working as per the prior specifications and guidelines. This timely checking of the system will help governance and increase the user's trust in the system. Another critical aspect of the Ex-ante accountability methods is to specify who will be directly or indirectly affected by the system. This is done using the impact assessment of the model through pre-trials. Different types of impact assessment methods are available based on the context requirements of the system. Some of the available impact assessments are Human Rights Impact Assessment (HRIA) [105], Privacy Impact Assessment (PIA) [163], Ethical Impact Assessment (EIA) [214], and Surveillance Impact Assessment (SIA) [215]. These methods are useful in laying out all the impacts of the AI system to its stakeholders. Kaminski and Malgieri [96] proposed an accountability and governance method in compliance with the GDPR by using the multi-layer explanations from the algorithmic impact assessment. They argued that one type of explanation does not satisfy all stakeholders; multi-layered explanations for assessment are needed to fulfill and foster individual rights and provide a central framework for systematic governance. All of these proposed methods deal with clearly defining algorithmic design and implementation specifications for better governance and accountability.

In-Ante methods. These methods deal with implementing accountability measures in the development lifecycle of the AI system itself and ensure that AI systems are developed as specified in the planning and design phase. These methods govern every step of the development to ensure that the resulting system will not cause any harm and that it is a fair and accurate system. This governance at every step of development is essential because once the system is deployed, it becomes challenging to track the causes of errors. Crawford [45] claimed that accountability of the AI system does not depend upon a single element but multiple elements like the choice of data, algorithms, weighting criteria, and objective functions, among others. So, it becomes imperative to govern the system 'end-to-end'. The first step is to ensure that the data used for training is diverse and does not have any bias [55, 132, 188]. The second step guarantees that an appropriate model has been chosen based on the problem requirements [65, 152]. This step deals with making sure that the model's accuracy satisfies some threshold for reliable decision making. It also deals with the careful usage of sensitive attributes and using different interpretability methods to assure that valuable features are used for the decision-making process [55]. The last step is to ensure that

Table 3. Accountability: Summarizes Proposed Methods Based on the Level of Human Involvement, and Each Method Is Labeled with Control Points That Can Be Used to Review Their Effectiveness in the AI Lifecycle, as Described in Figure 3

Level of Human Involvement	Control Points	Reference
Before-the-loop	1	[25, 26, 29, 96, 105, 134, 163, 174, 214, 215]
In-the-loop	2	[55, 55, 65, 132, 152, 188]
	2, 3	[159, 174]
Over-the-loop	4	[73, 111, 116, 119, 174, 197]
	1, 4	[157]

proper testing is done before deploying the system [116]. This step deals with the amount of decision information and explanations given to different stakeholders [132, 188] and specifying if-else conditions provides guidelines for how the system should perform and be used [220, 222].

However, other researchers [138] argued that opening the black box and providing total transparency to the public does not ensure accountability. They said that total transparency could lead to many adverse effects like loss of privacy, loss of trade secrets, and more attacks. There can be cases where total transparency can lead to less answerability. Because of these side effects, they proposed that total transparency for accountability should not be provided to the general public but only to the oversight agencies the public should trust [52]. Some researchers have also proposed a systematic framework to govern the development of an AI system. Raji et al. [159] proposed an internal auditing framework (SMACTR) for algorithms that create audit documents at every step of the development to keep track of all the decisions and evaluations. This technique leads to better controllability of the development process.

Post-Ante methods. Post-Ante methods deal with providing accountability measures after the model is deployed. Kroll et al. [116] proposed a legal framework to ensure that the deployed model works within the specified boundaries. This method makes use of technical tools to approximate evidence-based correctness for oversight. Some researchers suggested that collaboration between researchers, policymakers, and developers through auditing is needed for better accountability. LaBrie and Steinke [119] proposed an ethical algorithm audit that can be used for external validation of the system, helping the deployed system to be free from biases and errors and safe to use. Clavell et al. [73] proposed an algorithmic auditing technique to detect algorithmic bias in the system. They discussed how auditing techniques would help operationalized bias mitigation methods and lead to the development of fair algorithms. Sandvig et al. [174] proposed algorithmic auditing techniques for Internet platforms based on traditional field-based auditing techniques. They offered five auditing techniques—code audit, noninvasive audit, scraping audit, sock puppet audit, and collaborative audit to test different features and detect flaws in the algorithm. Another set of researchers [157] proposed a social contract method to monitor the working of algorithms based on agreements using society-in-the-loop. This method takes into account stakeholders with similar or conflicting interests and values to evaluate the system. Tutt [197] proposed a federal agency for algorithms that approves or disapproves algorithmic systems after performing proper pre-trials and ensuring that they are safe regarding their potential risks. This agency can also enforce guidelines for the use of algorithms, hence preventing their misuse.

Table 3 summarizes all of these accountability methods based on their involvement in the AI lifecycle. Each method is labeled with a control point, identifying the level to be applied and effectively evaluated.

3.3.3 Technical Challenges. Accountability of algorithms is an incremental process [18], requiring proper governance of the AI lifecycle and discussion between different stakeholders [144]. However, answering the question of who is responsible when the system is not working correctly

is challenging because multiple stakeholders are involved in the development process. Choosing which stakeholder negligence has led to the error is difficult. This is why proper accountability measures should be designed based on the application domain, as one policy cannot be applied to all domains. ISO [92] raised a need for developing context- and application-dependent policies for better governance. They explained their point using two main applications of AI: medical AI and AI-based recruiting systems. In medical AI, users of the system, such as doctors, can be held accountable for the harm caused by the system because they are domain experts in the field. Therefore, they should only use AI systems to assist their decision making, not for making decisions for them. Whereas in the recruiting system, we cannot hold system users responsible for the failures, as they may not know why the application got rejected. So, more research is needed for context-dependent accountability measures.

3.4 Privacy

AI-based decision-making systems analyze a vast amount of data to make decisions, where accuracy and performance depend upon the amount of data used for training. However, the considerable availability and usage of data can also have adverse effects. Private organizations, governments, or hackers could misuse our data, leading to harmful consequences. For example, a government misused citizens' personal data leading to inaccurate assessment of debt [24]. In addition, a social networking company collected and shared the personal information of 50 million users without their consent, which has been used to manipulate the presidential elections in the United States [34]. These examples show that it is vital to protect the privacy of the data to both avoid harmful consequences and increase the users' trust in the system. If the users know that the AI system they are using has taken appropriate measures to protect their identity and data, they tend to trust that system more.

3.4.1 Need for Privacy. With the increased availability of data, data breach incidents have also increased. Some examples of these breaches are when Equifax data breach exposed the personal data of millions of users [15] and when hackers gained access to 40 million credit and debit card details of Target customers [80]. These data breaches have exposed billions of pieces of sensitive information records and have led to potential abuse of the information. In addition, internal attacks or targeted attacks can cause data breaches. Such attacks can bring down the deployed system hence decreasing a user's trust in the system. Therefore, for trustworthy AI systems, it is crucial to protect the privacy of the system and the users. There can be different privacy threats to AI systems during the data collection, pre-processing, modeling, and implementation phases. In the data collection phase, privacy threats can be due to the data collected and its storage issues. Privacy can be compromised in the pre-processing and modeling phase if the AI system can re-identify sensitive information from non-sensitive data [190]. Another threat can be someone learning the internal functioning of the model by repeatedly querying the model [121]. Therefore, it is essential to protect the privacy of the data and the model to make AI systems trustworthy.

3.4.2 Proposed Solutions. Different methods have been proposed to ensure privacy. One way to ensure data privacy is through de-identification techniques that remove personal identifiers and their associations from the data [74]. These identifiers can be of two types: direct and indirect. Direct identifiers are the ones that are directly linked to a person's identity. In contrast, indirect identifiers can identify people's identity when linked with other information [74]. In the literature, various de-identification techniques are presented. Garfinkel [74] proposed data sampling and aggregation techniques for de-identification, where they represented the whole dataset through a sub-sample or summarized version that avoids releasing the entire dataset. Khalil and Ebner [107] proposed a masking technique to mask the values of sensitive attributes. These techniques are

called *suppression techniques* that increase the data's privacy but decrease the utility and quality. Another method proposed by the researchers is de-identification through pseudonymization [189]. This method removes the association between personal identifiers and adds an association between data subject characteristics and pseudonyms to ensure privacy. Researchers [221] have also discussed different encryption techniques like format-preserving [59] and order-preserving encryption [106] for de-identification.

Some researchers have proposed federated learning to provide privacy to the user data using the concept of collaborative learning where the raw data is not shared among the devices; instead, the model is shared after being trained on local devices [194]. Various federated learning models have been proposed. McMahan et al. [133] proposed a distributed learning method that uses data stored on mobile devices to compute a shared learning model, which prevents the need to centralize data storage providing data privacy as data is distributed over users' mobile devices. To preserve user-level differential privacy, Geyer et al. [76] proposed a framework that maintains the contribution of the user data in distributed learning. These federated learning models still have the risk of adversaries if the shared parameters have been exploited. To avoid this, Hao et al. [85] proposed a PEFL (Privacy Enhanced Federated Learning) approach, a non-interactive approach that can prevent data leaks when different entities interact. A detailed analysis of these methods is presented in the work of Li et al. [125].

Researchers have also proposed privacy measurement techniques to measure the risk of re-identification. One such method is the k-anonymity model [191], which deals with releasing data so that its ability to associate with other attributes using indirect identifiers is limited. This formal model limits the linkability to 1/k where k is at least the number of records present in the equivalence class for each attribute. However, this method has serious privacy problems [129], as the dataset released by this method can be easily attacked, and sensitive attributes can be accessed if background knowledge is available or if the sensitive attribute values lack diversity. To overcome these issues, L-diversity [129] was proposed, which is a refinement of the k-anonymity method. L-diversity adds the notion of intra-group diversity to the anonymization process for increased privacy. Another such formal model is the t-closeness method [124], which deals with datasets having an uneven or categorical distribution of the attributes. This method extends L-diversity by taking the distribution of sensitive attribute values into account. Regardless of the effectiveness of these methods, Rocher et al. [164] show that they can be insufficient. To overcome this, ISO [92] proposed that the risk of re-identification can only be managed through proper data agreements between the data-sharing parties.

Another concern about privacy is that once the data has been shared, it is challenging to delete or forget the data's online presence. The data used for AI system training cannot be easily removed, as these models memorize the data. There is a need for a framework that gives users total controllability over their data. For this, researchers have proposed an approach called *machine unlearning* [35], a system forgetting method that can be used to hide or forget sensitive user data and their lineages to enforce privacy. It uses the concept of creating and updating model summations to ignore the data. This type of method will provide users complete controllability over their data. Bourtoule et al. [22] also proposed an unlearning method called *SISA* (Shared, Isolated, Sliced, and Aggregated) training, which can expedite the unlearning process by decreasing the influence of sensitive data in the training phase, hence making the unlearning process easy and fast.

All of these methods can be used to ensure privacy in the decision-making process. Table 4 summarizes these methods and has categorized them based on their implementation at different AI lifecycle stages. For example, de-identification techniques are applied at the data pre-processing phase, and their effectiveness can be evaluated later at control point 2. In contrast, federated learning techniques are applied in the modeling phase and can be assessed at control point 3. Table 4

Table 4. Privacy: Summarizes Proposed Methods Based on Different Privacy-Preserving Approaches, and Each Method Is Labeled with the Level of Human Involvement and Control Points as Described in Figure 3

Level of Human Involvement	Control Points	Method Type	Reference
Before-the-loop	1	Data Agreements	[92]
In-the-loop	2	De-Identification Techniques	[59, 74, 106, 107, 189, 221]
	3	Federated Learning	[76, 85, 133, 194]
	3	Machine Unlearning	[22, 35]
Over-the-loop	4	Privacy Measurement Models	[124, 129, 191]

categorizes the proposed solutions based on their control points to help developers of the system carefully select the solution based on the requirements.

3.4.3 Technical Challenges. Some researchers proposed de-identification and federated learning techniques to ensure privacy, whereas others proposed using data agreements to enforce privacy. All of these methods have their pros and cons and should be implemented based on the application requirements. The selection of these techniques should be based on the trade-off between the performance and privacy-preserving overhead [39]. Despite the availability of these methods, the privacy of AI systems is still a significant challenge. There is a need for privacy laws that allow AI systems to benefit society and ensure privacy, but providing both at the same time can sometimes be difficult. To demonstrate this, Rosenquist [166] provided an excellent example. They stated that AI systems could be beneficial in rescuing exploited children through social media using facial recognition analysis of all the pictures posted on social media, which is impossible for humans to accomplish but still against the privacy laws of the individuals. So, there is a need for clear context-based privacy laws to define allowability and the conditions for that allowance.

3.5 Acceptance of AI

With the increase in the use of AI-based decision-making systems, it becomes essential to make them reliable and trustworthy. However, several failures in these systems have made them less reliable, which led to less acceptance and mistrust among the system users. There is a need for a mechanism that can increase the acceptance and trust for AI-based decision-making systems by carefully evaluating the system. Different researchers have proposed methods to increase acceptance and trust. Some researchers [146] presented various factors necessary for building trust, such as performance, type of task, application type, human component, and explainability of the system. Other researchers [49] described the importance of human involvement in increasing confidence in AI systems and making humans liable for their decisions. Another set of researchers [192] proposed the need for separate governance laws to increase the acceptance of these systems. All of these different proposals aim to increase the acceptance of AI systems among users through system evaluation. Some research has been done in this field to develop acceptance mechanisms. Before discussing the solutions, we will first discuss the need for such mechanisms.

3.5.1 Need for Acceptance of AI. Different users of the AI system can have different expectations. However, inadequate information and understanding of the system can lead to inflated expectations. When not fulfilled, these huge expectations can cause less acceptance of the systems. These bloated expectations can be related to the system's performance, usability, reliability, and fairness, among others. Some studies [151] show that the users of the AI system are still failing to fully embrace its uses, despite its numerous benefits. They found that the main reason for low acceptance was people's mistrust in the system for its lack of empathy and morality. Potential users wanted mechanisms to evaluate the system based on their trustworthy requirements and ethical

principles. That is why there is a need for a tool that can be used to assess these AI systems based on the users' requirements and expectations.

3.5.2 Proposed Solutions. Not much work has been accomplished in the field of Acceptance of AI. Different researchers have proposed various methods to increase the acceptance of AI among users. These methods are briefly described next.

Some researchers have suggested that one way to increase the acceptance of an AI system is by clearly understanding the expectations of different stakeholders and then designing the system accordingly. Researchers [113] showed that changing the model's focus can lead to change in the user acceptance of the AI system. For example, two users may expect that the system attains at least 50% accuracy but differ in their expectations of the type of falses—that is, wanting fewer false positives or false negatives. These different expectations from a diverse group of users will lead to different acceptance rates. Kocielnik et al. [113] explored various techniques to shape user expectations and maximize acceptance based on how the AI system is described, how well the user understands the system, their first impression, and so on. All of these factors play an essential role because users may perceive information differently, leading to different acceptance levels. The researchers showed how people accept false positives instead of false negatives for AI scheduling assistants because of the low recovery cost.

Some researchers have proposed theoretical models to analyze users' willingness to use AI systems. Gursoy et al. [82] proposed a theoretical model to study the user's willingness to use the AI device in the service delivery context. They said that the user's acceptance of the AI systems is based on their performance and effort expectations. Performance expectation is related to the accuracy and error rate of the model—that is, how much the technology will ease the task. In contrast, effort expectations deal with the amount of effort needed to use the system effectively. They proposed a three-level AI acceptance model. The first level deals with the user's general evaluation of the AI system. This involves social influence, motivation to use the system, and first-hand experience of it. The second level deals with calculating performance and effort expectancy from the attributes of the previous level. The last level is the outcome stage, where acceptance or rejection is decided based on previous level performance and effort expectancy. In this model, they showed how social influence from society and hedonic motivation positively impact performance expectancy.

Other researchers [204] have proposed a theoretical model called *Unified Theory of Acceptance and Use of Technology* (UTAUT 2) that can be used to predict user behavior toward new technology/AI systems. They have listed various extrinsic and intrinsic indicators that can influence user behavior. They showed how users' age, gender, experience, and habits could also affect the user acceptance and behavior toward the AI system. Wang et al. [208] proposed a theoretical model that analyzes the users' perceived need and affordability of upgrading to an AI system. This model deals with the willingness of the user to adopt new upgrades/technology. This willingness results from the trade-off between motivational factors (need, benefit, etc.) and non-motivational factors (cost, habit change, etc.). This model shows that demand and cost also play a vital role in accepting the AI system. Sohn and Kwon [184] compared various technology acceptance models that can be implemented in AI to increase users' acceptance and satisfaction.

Some researchers proposed that trust plays a significant role in accepting AI systems and is crucial for relationship building between different entities, whether it is a human-human relationship or a human-machine relationship. Trust building is a slow continuous process and is evidence based [8]. Different mechanisms have been proposed to calculate trust between various entities. Ruan et al. [168] proposed a trust mechanism to capture human-machine interactions, which can be used for trust building between users and AI systems. This mechanism considers history and

Table 5. Acceptance: Summarizes Different Acceptance Techniques Based on Their Level of Human Involvement Needed and Control Points in the AI Lifecycle Where These Methods Can Be Reviewed for Their Effectiveness as Described in Figure 3

Level of Human Involvement	Control Points	Reference
Before-the-loop	1	[82, 113, 208]
Before-the-loop, Over-the-loop	1, 4	[82, 204]
In-the-loop, Over-the-loop	3, 4	[103, 168, 203]

evidence to calculate trust and can help users carefully evaluate the AI systems. If the system has more frequent positive evidence of correct decision making, its trust will be high compared to the same type of system with less frequent evidence of proper decision making. This mechanism helps the users compare different AI systems. Some researchers [199–202] implemented this trust mechanism for an AI-based resource allocation system in FEW (Food, Energy, and Water), a high-stakes application because one wrong decision in resource allocation can drastically affect the lives of many farmers. Kaur et al. [104] proposed that trust can be context dependent and showed the potential for cases where algorithms perform better than humans and vice versa. For such cases, there is a need for human-machine collaboration. Kaur et al. [101] also showed how an AI-based fake user prediction system could use community knowledge to build trust. Other researchers [103, 203] proposed a trust-based acceptance metric to calculate the system's acceptance based on the explanations provided by it. All of these methods show how trust building through stakeholder involvement can help increase the acceptance of AI systems and describe the importance of trust building and acceptance of AI systems. Table 5 summarizes all of these acceptance methods based on their level of involvement in the AI lifecycle. It categorizes methods based on the control points to demonstrate different AI lifecycle phases where the method can be implemented and evaluated.

4 VERIFICATION AND VALIDATION

The trustworthy AI requirements, which are fairness, explainability, accountability, privacy, and acceptance, will make the AI systems ethical, increasing the users' trust in the system. Different methods have been proposed to satisfy additional trustworthy AI requirements. Some methods satisfy more than one trustworthy AI requirement. However, it can be challenging to meet all trustworthy requirements simultaneously without compromising the system's performance. There is a need for a trade-off mechanism that can prioritize different trustworthy requirements and accuracy based on the application needs. Some AI applications may require maximizing trustworthy requirements based on a given accuracy constraint, whereas others can require maximizing accuracy based on the given trustworthy requirements constraint. To ease this process, we have summarized different proposed solutions based on the trustworthy requirements they satisfy and their effect on the system's accuracy. Table 6 contains this comparison of different solutions based on trustworthy AI requirements. This helps the designers and developers of the system to carefully select the appropriate, trustworthy AI solutions based on their needs. It also provides insight to the researchers in this field about how different methods present in literature satisfy multiple trustworthy AI requirements and affect the performance of the AI systems.

After carefully designing and developing the AI system, it is crucial to test and validate the system to satisfy all of the defined trustworthy AI requirements. This plays a vital role in the trustworthiness of the AI system. Different approaches have been described to verify and validate AI systems, and applications might have varying needs. In this section, we provide an overview of these techniques.

Table 6. Comparison of Different Proposed Methods That Satisfy Various Trustworthy AI Requirements and How These Methods Affect the Accuracy of the AI System

Effect on Accuracy	Trustworthy Requirements	Reference
No effect	F, E, A	[123, 170]
	F, A	[22, 171, 226]
	F, E	[14, 179]
	E, A	[75, 108, 186, 206]
	F, E, A, P, Ac	[92]
	F, E, A, Ac	[134]
	F, Ac	[26]
	A, P, Ac	[163, 214]
	E, A	[55, 132]
	A, Ac	[25, 29, 105, 113, 157, 159, 168, 197]
	F, A, Ac	[111, 174]
	Ac	[204]
Decrease	F	[10, 27, 33, 38, 97, 128, 172]
	F, P	[63, 91, 169, 189]
	F, A	[156]
Increase	F	[135]
	F, P	[86]

F, Fairness; E, Explainability; A, Accountability; P, Privacy; Ac, Acceptance.

AI systems contain both deterministic and non-deterministic components. Both components need to be evaluated for the verification and validation of the system. Deterministic components are the ones whose behavior is fully known and can be clearly predicted. Therefore, these components can be tested using the traditional approaches of software testing. However, for non-deterministic components, which is the AI component, conventional techniques cannot be used because there is no clear standard available to specify its requirements, also known as the oracle problem [92]. The oracle problem is defined as when there is no precise mechanism available to test whether the system is working correctly, therefore requiring extensive human involvement to check the output generated by the system [210]. AI and machine learning systems suffer from this problem, as it is challenging to create an oracle to check the output without human involvement. For example, in face recognition algorithms, data on which the system is trained and tested is labeled by humans, thus limiting the scope of testing as the sample size and diversity of the testing dataset depends on human efforts [178]. To overcome this issue, different approaches and methods have been proposed, as discussed next:

- *Metamorphic testing:* This type of testing is used to avoid the oracle problem. It is based on the input and output relationship of the system. This method tests the system by testing the input-output relationship for multiple iterations of the system. This testing approach is based on a simple principle that if the correct output for the input is not known, we can test the system based on the outputs of multiple related inputs [41]. Several researchers have used this technique to test AI systems. Lindvall et al. [126] used it to test the control software for autonomous drones. Xie et al. [217] have tested different classifiers using this technique. In these cases, metamorphic testing detected unexpected faults and assumptions in the AI system.
- *Expert panels:* Expert panels can be used when traditional testing methods are not possible. This technique can be helpful when the AI system is built to assist or replace experts [92]. This independent panel of experts is responsible for providing possible diagnoses and recommendations for the outputs of the AI system [185]. Knauf et al. [112] used this method

to test a rule-based system where experts are provided with carefully designed test cases, and the decision made by an expert is compared to those created by both the other experts and the AI system. The problem with this type of method lies in resolving a disagreement between experts.

- *Benchmarking:* Benchmarking is a technique that is used to test, measure, and compare the performance of the AI system on publicly available carefully designed datasets [92]. Different researchers and government agencies are working to create benchmarks for various AI applications. Jiang et al. [94] developed a benchmark suite HPCAI500 for high-performance scientific computing (HPC) AI systems. Ngan et al. [145] designed the face recognition vendor test FRVT to compare and assess the performance of automated gender classification algorithms. These advancements show the importance of creating benchmarks for AI systems. However, AI systems are diverse, so it is challenging to develop a single benchmark. To overcome this issue, the TPC (the Transaction Processing Performance Council) formed a group known as TPC-AI in 2018 to create benchmarks for AI-based workloads [143]. These benchmarks will help to better test and validate AI systems.
- *Field trials:* The field trials technique is used to test the performance and durability of the system in a real operating environment. It is a valuable technique because it shows how the actual users will interact with the AI system. This testing method is used when the testing environment is entirely different from the existing environment [92]. Field trials are also an effective way to check the AI system's acceptability by real users. Some domains have used field trials to test the behavior and performance of AI systems under different conditions. Bundesamt [28] used field trials to test facial recognition systems. The UK is using field trials to test self-driving cars [70]. These examples show that field trials are a valuable technique to test AI systems under uncertain real-world conditions.
- *Testing in a simulated environment:* Testing in a simulated environment is beneficial when the AI system is designed to perform physical actions in the environment [92]. This type of testing can be performed on robots and drones with AI systems embedded in them. In simulated testing, a controlled environment is used to evaluate the system's performance under different conditions.
- *Comparison to human intelligence:* This type of test is used to evaluate AI systems designed to perform tasks traditionally done by humans or that need human cognitive abilities. Users, regulators, or policymakers of the AI system typically perform these tests to evaluate performance. This type of test is helpful when the decisions made by the AI system are compared to those of a trained/licensed professional in that field [92]. If the system performs as well as the trained human professionals, this can create a sense of user trust toward the system. In these tests, carefully designed data samples are used to compare the decisions of humans and AI systems.

All of these mechanisms test the AI systems from different aspects and requirements. Many AI systems are probabilistic, which results in non-reproducibility [92], making it difficult to test every aspect of these systems using existing practices. So, there is a need to carefully develop standards for the verification and validation of AI systems.

5 DISCUSSION AND FUTURE DIRECTIONS

In this review, we have tried to cover a broad spectrum of methods proposed to make AI trustworthy and answer the research questions raised in Section 1:

- *R1:* The requirements to make AI systems trustworthy are making them lawful, ethical, and robust. This means that the AI development, deployment, and use should follow all

applicable laws and regulations; respect and follow the humans' ethical principles and guidelines such as fairness, explainability, accountability, privacy, and acceptance; and be technically robust and reliable.

- *R2:* To govern the operation of AI systems, some guidelines have been developed [43, 92]. However, there is still an implementation gap between the research and practice. So, there is a need to establish policies and standards to enforce these guidelines and existing laws into practice.
- *R3:* Human involvement is essential in this changing era of AI because this new era is moving toward collaborative thinking [51], which uses humans' cognitive ability and machines' exceptional computing power to reach the best decision making. AI systems are being used in various high-stakes applications, where the consequences of failures are hazardous. So, to make these systems safe, reliable, and trustworthy, humans are needed to develop efficient algorithms, set limits for performance, flag and correct errors raised by the system, override wrong decisions, and improve the system's performance.
- *R4:* To make AI decisions acceptable and to increase the usability and trust of the system, users should clearly understand its usability, performance, and limitations. A proper evaluation mechanism is needed to evaluate the system based on trustworthy requirements and users' expectations to prevent any bloated expectations.

The answers to these research questions provide an important stepping stone in the research of trustworthy AI. However, it is impossible to cover all critical aspects of trustworthy AI. This review offers an overview of the current stage of research in trustworthy and ethical AI. In the following, we present some of the future directions and open problems we found in this field of trustworthy AI:

- As AI is becoming an essential part of today's digital economy, developing *standards and policies* to govern and utilize these systems to their full potential has become crucial. Standardization of AI will lead to faster technology transfer, interoperability, security, and reliability [216]. Therefore, developing new standards is necessary, but AI systems should also abide by the existing laws and regulations based on their use cases [92]. Thus, it is imperative to develop not only AI standards but also policies to enforce the standards.
- AI is being deployed in a various range of applications. To utilize the full potential of AI, especially in safety-critical applications, there is a need to make these algorithms safe and trustworthy. The safety and trustworthiness of AI systems mainly depend on their ability to provide explanations to different stakeholders—designers, developers, users, domain experts, and policymakers [12]. To satisfy the need of all the stakeholders, there is a need for *multidisciplinary research* involving data science, computer science, sociology, economics, and law to build and implement AI applications. This type of research will bring experience and expertise from different backgrounds to make AI safe and trustworthy.
- Different stakeholders involved with the AI system can have varying expectations for the capabilities of the system. These expectations affect their acceptance rate and trust in the system [158]. To avoid inflated or low expectations, there is a need for an *expectation management framework*. Some work has been done in this field that deals with post-development expectation management, which captures how different factors like information about the system, reasoning and understanding of the system, and first-hand experience of the system can affect the system's acceptance [56, 113]. There is a need for such an expectation management framework from the start of developing the AI system. This framework can help clarify the designers' expectations about what the users expect from the system to design and develop the system accordingly [92].

- Some studies [151] have shown that the potential users of the AI system are still failing to use and accept them, despite their numerous benefits. They found out that it is because the potential users want some measurement mechanism that can quantitatively evaluate the trustworthy AI requirements of the system. As a result, some theoretical models [204], trust models [168], and human involvement methods [51] have been proposed to increase the user acceptance of these systems. But still, there is a lack of *measurement mechanisms* that can test and evaluate the users' acceptance based on carefully designed testbeds to measure the effectiveness of trustworthy AI requirements and their effect on the users.

6 CONCLUSION

This review has revolved around trustworthy AI, an important research field in the AI ecosystem. We have elaborated on this topic by first discussing the need and importance of trust development in AI systems. We then discussed and reviewed different requirements to make AI trustworthy and their corresponding proposed methods. Last, we discussed various testing techniques to verify and validate the AI systems based on trustworthy requirements. This review provides a global taxonomy of the proposed methods and recent developments in the field of trustworthy AI. In this review, the proposed methods were classified based on their level of implementation at different levels of the AI lifecycle to provide a clear picture to the readers. While designing and developing this review, we tried to focus on the central idea that AI systems are there to empower and complement humans, not to replace them. That is why we also shed some light on how humans can be involved at different AI lifecycle stages to make the system trustworthy.

In this review, we concluded that the AI systems' benefits could be impeded if members of the society do not trust the system to perform as intended and not cause any harm to the users or the society at large. Trust in the AI system depends on fairness, explainability, accountability, privacy, and user acceptance. To ensure that these requirements are met, there is a need for proper mechanisms, standards, and legal frameworks to govern the development and working of AI systems. To summarize in one line—a standardization/legal framework is needed to govern AI systems, which will increase the users' trust in the system, leading to greater social acceptance of these systems.

REFERENCES

[1] Peter Achinstein. 1983. *The Nature of Explanation*. Oxford University Press on Demand.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Aniya Agarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2018. Automated test generation to detect individual discrimination in AI models. *arXiv preprint arXiv:1809.03260* (2018). https://arxiv.org/abs/1809.03260.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*, May 23, 2016.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[6] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv e-prints* (2019), arXiv–1909. https://arxiv.org/abs/1909.03012.

[7] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 1770–1780.

[8] Sulin Ba. 2001. Establishing online trust through a community responsibility system. *Decision Support Systems* 31, 3 (2001), 323–336.

[9] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10, 7 (2015), e0130140.

[10] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *Proceedings of the International Conference on Machine Learning*. 405–413.
[11] Edelman Trust Barometer. 2019. Edelman Trust Barometer Global Report. Retrieved November 2, 2021 from https://www.edelman.com/sites/g/files/aatuss191/files/2019-02/2019_Edelman_Trust_Barometer_Global_Report.pdf.
[12] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. 2020. Flexible and context-specific AI explainability: A multidisciplinary approach. Available at SSRN 3559477 (2020).
[13] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017). https://arxiv.org/pdf/1707.00044.pdf.
[14] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4–5 (2019), Article 4, 15 pages.
[15] Hal Berghel. 2017. Equifax and the latest round of identity theft roulette. *Computer* 50, 12 (2017), 72–76.
[16] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017). https://arxiv.org/abs/1706.02409.
[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.
[18] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*. 1–14.
[19] Emily Black, Samuel Yeom, and Matt Fredrikson. 2020. FlipTest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 111–121.
[20] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity*. Technical Report. Upturn.
[21] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.
[22] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP'21)*. IEEE, Los Alamitos, CA, 141–159.
[23] Jon Boyens, Celia Paulsen, Rama Moorthy, Nadya Bartol, and Stephanie A. Shankles. 2015. Supply chain risk management practices for federal information systems and organizations. *NIST Special Publication* 800, 161 (2015), 32.
[24] Valerie Braithwaite. 2020. Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. *Australian Journal of Social Issues* 55, 3 (2020), 242–259.
[25] Kiel Brennan-Marquez. 2017. Plausible cause: Explanatory standards in the age of powerful machines. *Vanderbilt Law Review* 70 (2017), 1249.
[26] Dennis Broeders, Erik Schrijvers, Bart van der Sloot, Rosamunde van Brakel, Josta de Hoog, and Ernst Hirsch Ballin. 2017. Big data and security policies: Towards a framework for regulating the phases of analytics and use of big data. *Computer Law & Security Review* 33, 3 (2017), 309–323.
[27] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the International Conference on Machine Learning*. 803–811.
[28] Fursind Bundesamt. 2004. *Study: "An Investigation into the Performance of Facial Recognition Systems Relative to Their Planned Use in Photo Identification Documents–BioP I."* Bundesamt fur Sicherheit in der Informationstechnik.
[29] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. 169–178.
[30] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 77–91.
[31] B. Burke, D. Cearley, N. Jones, D. Smith, A. Chandrasekaran, C. K. Lu, and K. Panetta. 2019. Gartner Top 10 Strategic Technology Trends for 2020-Smarter with Gartner. Retrieved November 2, 2021 from https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020/.
[32] Ewen Callaway. 2021. DeepMind's AI predicts structures for a vast trove of proteins. *Nature* 595, 7869 (2021), 635–635.
[33] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[34] Chiara Campione. 2020. The Dark Nudge Era: Cambridge Analytica, Digital Manipulation in Politics, and the Fragmentation of Society. Bachelor's Thesis. Luiss Guido Carli.
[35] Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy*. IEEE, Los Alamitos, CA, 463–480.
[36] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. 2018. Visualizing the feature importance for black box models. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 655–670.
[37] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.
[38] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2016. How to be fair and diverse? *arXiv preprint arXiv:1610.07183* (2016).
[39] Huili Chen, Siam Umar Hussain, Fabian Boemer, Emmanuel Stapf, Ahmad Reza Sadeghi, Farinaz Koushanfar, and Rosario Cammarota. 2020. Developing privacy-preserving AI systems: The lessons learned. In *Proceedings of the 2020 57th ACM/IEEE Design Automation Conference (DAC'20)*. IEEE, Los Alamitos, CA, 1–4.
[40] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.
[41] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys* 51, 1 (2018), 1–27.
[42] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 134–148.
[43] European Commission. 2020. *White Paper on Artificial Intelligence—A European Approach to Excellence and Trust*. European Commission.
[44] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.
[45] Kate Crawford. 2016. Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41, 1 (2016), 77–92.
[46] Kate Crawford. 2021. *The Atlas of AI*. Yale University Press.
[47] Bruno Silveira Cruz and Murillo de Oliveira Dias. 2020. Crashed Boeing 737-MAX: Fatalities or malpractice? *GSJ* 8, 1 (2020), 2615–2624.
[48] Angela Daly, S. Kate Devitt, and Monique Mann. 2021. AI ethics needs good data. *arXiv preprint arXiv:2102.07333* (2021). https://arxiv.org/ftp/arxiv/papers/2102/2102.07333.pdf.
[49] M. D. Danny Tobey. 2019. Explainability: Where AI and Liability Meet: Actualités: DLA Piper Global Law Firm. Retrieved November 2, 2021 from https://www.dlapiper.com/fr/france/insights/publications/2019/02/explainability-where-ai-and-liability-meet/.
[50] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Available at https://www.reuters.com.
[51] Paul R. Daugherty and H. James Wilson. 2018. *Human+ Machine: Reimagining Work in the Age of AI*. Harvard Business Press.
[52] Paul B. De Laat. 2018. Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology* 31, 4 (2018), 525–541.
[53] S. Kate Devitt. 2018. Trustworthiness of autonomous systems. In *Foundations of Trusted Autonomy*. Springer, Cham, Switzerland, 161–184.
[54] Virginia Dignum. 2017. Responsible artificial intelligence: Designing AI for human values. *ICT Discoveries* 1 (2017), 1–8.
[55] James E. Dobson. 2015. Can an algorithm be disturbed? Machine learning, intrinsic criticism, and the digital humanities. *College Literature* 42, 4 (2015), 543–564.
[56] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 Chi Conference on Human Factors in Computing Systems*. 278–288.
[57] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.
[58] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 119–133.
[59] Morris Dworkin. 2016. Recommendation for block cipher modes of operation: Methods for format-preserving encryption. *NIST Special Publication* 800 (2016), 38G.

[60] European Commission. 2018. Ethics Guidelines for Trustworthy AI. Retrieved November 2, 2021 from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[61] Anthony Elliott. 2019. *The Culture of AI: Everyday Life and the Digital Revolution*. Routledge.
[62] Wenjuan Fan, Jingnan Liu, Shuwan Zhu, and Panos M. Pardalos. 2020. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research* 294, 1 (2020), 567–592.
[63] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.
[64] Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair AI: Challenges and opportunities. *Business & Information Systems Engineering* 62, 1 (2020), 1–7.
[65] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*. 2962–2970.
[66] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489* (2018), 237–246.
[67] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. 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. *Federal Probation* 80 (2016), 38.
[68] Luciano Floridi and Josh Cowls. 2019. A unified framework of five principles for AI in society. *HDSR* 1.1 (2019).
[69] Luciano Floridi, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. 2020. How to design AI for social good: Seven essential factors. *Science and Engineering Ethics* 26, 3 (2020), 1771–1796.
[70] Department for Transport (UK). 2015. *The Pathway to Driverless Cars: A Code of Practice for Testing*. Department for Transport (UK).
[71] Maria Jose Gacto, Rafael Alcalá, and Francisco Herrera. 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences* 181, 20 (2011), 4340–4360.
[72] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).
[73] Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. 2020. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 265–271.
[74] Simson L. Garfinkel. 2015. *De-Identification of Personal Information*. National Institute of Standards and Technology.
[75] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018). https://arxiv.org/abs/1803.09010.
[76] Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557* (2017).
[77] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*. 9277–9286.
[78] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine* 38, 3 (2017), 50–57.
[79] Mark Granovetter. 2018. Economic action and social structure: The problem of embeddedness. In *The Sociology of Economic Life*. Routledge, 22–45.
[80] Claire Greene and Joanna Stavins. 2017. Did the target data breach change consumer assessments of payment card security? *Journal of Payments Strategy & Systems* 11, 2 (2017), 121–133.
[81] David Gunning. 2017. *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency.
[82] Dogan Gursoy, Oscar Hengxuan Chi, Lu Lu, and Robin Nunkoo. 2019. Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management* 49 (2019), 157–169.
[83] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30, 1 (2020), 99–120.
[84] Tameru Hailesilassie. 2016. Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv:1610.05267* (2016).
[85] Meng Hao, Hongwei Li, Xizhao Luo, Guowen Xu, Haomiao Yang, and Sen Liu. 2019. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. *IEEE Transactions on Industrial Informatics* 16, 10 (2019), 6532–6542.
[86] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[87] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87 (2014), 96–110.

[88] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015). https://arxiv.org/abs/1503.02531.

[89] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[90] Sarah Holland, Ahmed Hosny, and Sarah Newman. 2020. The dataset nutrition label. *arXiv preprint arXiv:1805.03677[cs.DB]* (2020).

[91] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and fair classification. In *Proceedings of the International Conference on Machine Learning*. 2879–2890.

[92] ISO 24028:2020. 2020. *Information Technology–Artificial Intelligence–Overview of Trustworthiness in Artificial Intelligence*. Standard. International Organization for Standardization.

[93] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037* (2018). https://arxiv.org/abs/1809.08037.

[94] Zihan Jiang, Wanling Gao, Lei Wang, Xingwang Xiong, Yuchen Zhang, Xu Wen, Chunjie Luo, et al. 2018. HPC AI500: A benchmark suite for HPC AI systems. In *Proceedings of the International Symposium on Benchmarking, Measuring, and Optimization*. 10–22.

[95] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[96] Margot E. Kaminski and Gianclaudio Malgieri. 2020. Multi-layered explanations from algorithmic impact assessments in the GDPR. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 68–79.

[97] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[98] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE, Los Alamitos, CA, 869–874.

[99] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 35–50.

[100] Andreas Kaplan and Michael Haenlein. 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons* 62, 1 (2019), 15–25.

[101] Davinder Kaur, Suleyman Uslu, and Arjan Durresi. 2019. Trust-based security mechanism for detecting clusters of fake users in social networks. In *Proceedings of the Workshops of the International Conference on Advanced Information Networking and Applications*. 641–650.

[102] Davinder Kaur, Suleyman Uslu, and Arjan Durresi. 2020. Requirements for trustworthy artificial intelligence—A review. In *Proceedings of the International Conference on Network-Based Information Systems*. 105–115.

[103] Davinder Kaur, Suleyman Uslu, Arjan Durresi, Sunil Badve, and Murat Dundar. 2021. Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems. In *Proceedings of the International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS'21)*.

[104] Davinder Kaur, Suleyman Uslu, Arjan Durresi, George Mohler, and Jeremy G. Carter. 2020. Trust-based human-machine collaboration mechanism for predicting crimes. In *Proceedings of the International Conference on Advanced Information Networking and Applications*. 603–616.

[105] Deanna Kemp and Frank Vanclay. 2013. Human rights and impact assessment: Clarifying the connections in practice. *Impact Assessment and Project Appraisal* 31, 2 (2013), 86–96.

[106] Florian Kerschbaum. 2015. Frequency-hiding order-preserving encryption. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 656–667.

[107] Mohammad Khalil and Martin Ebner. 2016. De-identification in learning analytics. *Journal of Learning Analytics* 3, 1 (2016), 129–138.

[108] Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*. 2280–2288.

[109] Been Kim, Cynthia Rudin, and Julie A. Shah. 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*. 1952–1960.

[110] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the International Conference on Machine Learning*. 2668–2677.

[111] Pauline T. Kim. 2017. Auditing algorithms for discrimination. *University of Pennsylvania Law Review Online* 166 (2017), 189.
[112] Rainer Knauf, Avelino J. Gonzalez, and Thomas Abel. 2002. A framework for validation of rule-based systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 32, 3 (2002), 281–295.
[113] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[114] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*. 1885–1894.
[115] Puneet Kohli and Anjali Chadha. 2019. Enabling pedestrian safety using computer vision techniques: A case study of the 2018 Uber Inc. self-driving car crash. In *Proceedings of the Future of Information and Communication Conference*. 261–279.
[116] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review* 165 (2016), 633.
[117] Abhishek Kumar, Tristan Braud, Sasu Tarkoma, and Pan Hui. 2020. Trustworthy AI in the age of pervasive computing and big data. In *Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops'20)*. IEEE, Los Alamitos, CA, 1–6.
[118] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[119] Ryan C. LaBrie and Gerhard Steinke. 2019. Towards a framework for ethical audits of AI algorithms. In *Proceedings of the Conference on Data Science and Analytics for Decision Support*.
[120] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1675–1684.
[121] Taesung Lee, Ian M. Molloy, and Dong Su. 2019. Protecting cognitive systems from model stealing attacks. US Patent App. 15/714,514.
[122] Shane Legg and Marcus Hutter. 2007. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications* 157 (2007), 17.
[123] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.
[124] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy beyond k-anonymity and L-diversity. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering*. IEEE, Los Alamitos, CA, 106–115.
[125] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60.
[126] Mikael Lindvall, Dharmalingam Ganesan, Ragnar Árdal, and Robert E. Wiegand. 2015. Metamorphic model-based testing applied on NASA DAT—An experience report. In *Proceedings of the 2015 IEEE/ACM 37th International Conference on Software Engineering*, Vol. 2. IEEE, Los Alamitos, CA, 129–138.
[127] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.
[128] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. *k*-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 502–510.
[129] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *l*-Diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 3–es.
[130] Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.
[131] Bernard Marr. 2018. Is artificial intelligence dangerous? 6 AI risks everyone should know about. *Forbes* (2018).
[132] Kirsten Martin. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160, 4 (2019), 835–850.
[133] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*. PMLR, 1273–1282.
[134] Dan McQuillan. 2018. People's councils for ethical machine learning. *Social Media+ Society* 4, 2 (2018), 2056305118768303.
[135] Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Debiasing community detection: The importance of lowly connected nodes. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'19)*. IEEE, Los Alamitos, CA, 509–512.

[136] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 1–35.
[137] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 107–118.
[138] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 735–746.
[139] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[140] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.
[141] Jethro Mullen. 2015. Google Rushes to Fix Software That Served Up Racial Slur. Retrieved November 2, 2021 from https://www.cnn.com/2015/07/02/tech/google-image-recognition-gorillas-tag/.
[142] Patrick M. Murphy and Michael J. Pazzani. 1991. ID2-of-3: Constructive induction of M-of-N concepts for discriminators in decision trees. In *Machine Learning Proceedings 1991*. Elsevier, 183–187.
[143] Raghunath Nambiar. 2018. Towards an industry standard for benchmarking artificial intelligence systems. In *Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE'18)*. IEEE, Los Alamitos, CA, 1679–1680.
[144] Daniel Neyland. 2016. Bearing account-able witness to the ethical algorithmic system. *Science, Technology, & Human Values* 41, 1 (2016), 50–76.
[145] Mei Ngan, Patrick J. Grother, and Mei Ngan. 2015. *Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms*. U.S. Department of Commerce, National Institute of Standards and Technology.
[146] Claire Nicodeme. 2020. 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'20)*. IEEE, Los Alamitos, CA, 20–23.
[147] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[148] National Institute of Standards and Technology. 2021. NIST Proposes Method for Evaluating User Trust in Artificial Intelligence Systems. Retrieved November 2, 2021 from https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence-systems.
[149] U.S. Government Accountability Office. 2021. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. Retrieved November 2, 2021 from https://www.gao.gov/products/gao-21-519sp.
[150] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.
[151] Amy L. Ostrom, Darima Fotheringham, and Mary Jo Bitner. 2019. Customer acceptance of AI in service encounters: Understanding antecedents and consequences. In *Handbook of Service Science, Volume II*. Springer, 77–103.
[152] David J. Pauleen, David Rooney, and Ali Intezari. 2017. Big data, little wisdom: Trouble brewing? Ethical implications for the information systems discipline. *Social Epistemology* 31, 4 (2017), 400–416.
[153] Petra Perner. 2011. How to interpret decision trees? In *Proceedings of the Industrial Conference on Data Mining*. 40–55.
[154] Sundar Pichai. 2018. AI at Google: Our principles. *The Keyword*, June 7, 2018.
[155] Stefan Poier. 2020. Clean and green—The volkswagen emissions scandal: Failure of corporate governance? *Problemy Ekorozwoju* 15, 2 (2020) 33–39.
[156] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*. 677–688.
[157] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
[158] Eeva Raita and Antti Oulasvirta. 2011. Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers* 23, 4 (2011), 363–371.
[159] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.
[160] Chris Reed. 2018. How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170360.
[161] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[162] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *Proceedings of the IUI Workshops*.
[163] Stuart Ritchie. 2017. Privacy impact assessment System and associated methods. US Patent App. 15/459,909.
[164] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10, 1 (2019), 1–9.
[165] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing bias in AI. In *Companion Proceedings of the 2019 World Wide Web Conference*. 539–544.
[166] Matthew Rosenquist. 2020. There Is No Easy Fix to AI Privacy Problems. Retrieved November 2, 2021 from https://www.helpnetsecurity.com/2020/01/23/ai-privacy-problems/.
[167] Julian B. Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of Personality* 35, 4 (1967), 651–665.
[168] Yefeng Ruan, Ping Zhang, Lina Alfantoukh, and Arjan Durresi. 2017. Measurement theory-based trust management framework for online social communities. *ACM Transactions on Internet Technology* 17, 2 (2017), 1–24.
[169] Salvatore Ruggieri. 2014. Using t-closeness anonymity to control for non-discrimination. *Transactions on Data Privacy* 7, 2 (2014), 99–129.
[170] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. 2020. Radioactive data: Tracing through training. In *Proceedings of the International Conference on Machine Learning*. 8326–8335.
[171] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018). https://arxiv.org/abs/1811.05577.
[172] Samira Samadi, Uthaipon Tantipongpipat, Jamie H. Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*. 10976–10987.
[173] Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 5–22.
[174] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on Internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* 22 (2014), 1–23.
[175] Lindsay Sanneman and Julie A. Shah. 2020. A situation awareness-based framework for design and evaluation of explainable AI. In *Proceedings of the International Workshop on Explainable, Transparent, Autonomous Agents and Multi-Agent Systems*. 94–110.
[176] Fernando P. Santos, Francisco C. Santos, Ana Paiva, and Jorge M. Pacheco. 2015. Evolutionary dynamics of group fairness. *Journal of Theoretical Biology* 378 (2015), 96–102.
[177] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[178] Sergio Segura, Dave Towey, Zhi Quan Zhou, and Tsong Yueh Chen. 2018. Metamorphic testing: Testing the untestable. *IEEE Software* 37, 3 (2018), 46–53.
[179] Lei Shi, Furu Wei, Shixia Liu, Li Tan, Xiaoxiao Lian, and Michelle X. Zhou. 2010. Understanding text corpora with multiple facets. In *Proceedings of the 2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, Los Alamitos, CA, 99–106.
[180] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
[181] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.
[182] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469* (2016). https://arxiv.org/abs/1611.05469.
[183] Nathalie A. Smuha. 2019. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* 20, 4 (2019), 97–106.
[184] Kwonsang Sohn and Ohbyung Kwon. 2020. Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. *Telematics and Informatics* 47 (2020), 101324.
[185] Richard S. Sojda. 2007. Empirical evaluation of decision support systems: Needs, definitions, potential methods, and an example pertaining to waterfowl management. *Environmental Modelling & Software* 22, 2 (2007), 269–277.
[186] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.
[187] Biplav Srivastava and Francesca Rossi. 2018. Towards composable bias rating of AI services. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 284–289.

[188] Bernd Carsten Stahl and David Wright. 2018. Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy* 16, 3 (2018), 26–33.
[189] Sophie Stalla-Bourdillon and Alison Knight. 2016. Anonymous data v. personal data-false debate: An EU perspective on anonymization, pseudonymization and personal data. *Wisconsin International Law Journal* 34 (2016), 284.
[190] Du Su, Hieu Tri Huynh, Ziao Chen, Yi Lu, and Wenmiao Lu. 2020. Re-identification attack to privacy-preserving data analysis with noisy sample-mean. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1045–1053.
[191] Latanya Sweeney. 2002. *k*-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
[192] Andreas Theodorou and Virginia Dignum. 2020. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence* 2, 1 (2020), 10–12.
[193] Mike Thomas. 2019. 6 Dangerous Risks of Artificial Intelligence. Retrieved November 2, 2021 from https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence.
[194] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. 2019. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*. 1–11.
[195] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.
[196] Ryan Turner. 2016. A model explanation system. In *Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP'16)*. IEEE, Los Alamitos, CA, 1–6.
[197] Andrew Tutt. 2017. An FDA for algorithms. *Administrative Law Review* 69 (2017), 83.
[198] UNI Global Union. 2017. *Top 10 Principles for Ethical Artificial Intelligence*. UNI Global Union, Nyon, Switzerland.
[199] Suleyman Uslu, Davinder Kaur, Samuel J. Rivera, Arjan Durresi, and Meghna Babbar-Sebens. 2019. Decision support system using trust planning among food-energy-water actors. In *Proceedings of the International Conference on Advanced Information Networking and Applications*. 1169–1180.
[200] Suleyman Uslu, Davinder Kaur, Samuel J. Rivera, Arjan Durresi, and Meghna Babbar-Sebens. 2019. Trust-based game-theoretical decision making for food-energy-water management. In *Proceedings of the International Conference on Broadband and Wireless Computing, Communication, and Applications*. 125–136.
[201] Suleyman Uslu, Davinder Kaur, Samuel J. Rivera, Arjan Durresi, and Meghna Babbar-Sebens. 2020. Trust-based decision making for food-energy-water actors. In *Proceedings of the International Conference on Advanced Information Networking and Applications*. 591–602.
[202] Suleyman Uslu, Davinder Kaur, Samuel J. Rivera, Arjan Durresi, Meghna Babbar-Sebens, and Jenna H. Tilt. 2020. Control theoretical modeling of trust-based decision making in food-energy-water management. In *Proceedings of the Conference on Complex, Intelligent, and Software Intensive Systems*. 97–107.
[203] Suleyman Uslu, Davinder Kaur, Samuel J. Rivera, Arjan Durresi, Mimoza Durresi, and Meghna Babbar-Sebens. 2021. Trustworthy acceptance: A new metric for trustworthy artificial intelligence used in decision making in food-energy-water sectors. In *Proceedings of the 35th International Conference on Advanced Information Networking and Applications (AINA'21)*. 208–219.
[204] Viswanath Venkatesh, James Y. L. Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly* 36, 1 (2012), 157–178.
[205] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare'18)*. IEEE, Los Alamitos, CA, 1–7.
[206] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology* 31 (2017), 841.
[207] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Available at SSRN (2020).
[208] Yu-Yin Wang, Yi-Shun Wang, and Tung-Ching Lin. 2018. Developing and validating a technology upgrade model. *International Journal of Information Management* 38, 1 (2018), 7–26.
[209] Soeren H. Welling, Hanne H. F. Refsgaard, Per B. Brockhoff, and Line H. Clemmensen. 2016. Forest floor visualizations of random forests. *arXiv preprint arXiv:1605.09196* (2016). https://arxiv.org/abs/1605.09196
[210] Elaine J. Weyuker. 1982. On testing non-testable programs. *Computer Journal* 25, 4 (1982), 465–470.
[211] Maranke Wieringa. 2020. 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*. 1–18.
[212] Oliver E. Williamson. 1993. Calculativeness, trust, and economic organization. *Journal of Law and Economics* 36, 1, Part 2 (1993), 453–486.
[213] H. James Wilson and Paul R. Daugherty. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review* 96, 4 (2018), 114–123.

[214] David Wright. 2011. A framework for the ethical impact assessment of information technology. *Ethics and Information Technology* 13, 3 (2011), 199–226.
[215] David Wright, Michael Friedewald, and Raphaël Gellert. 2015. Developing and testing a surveillance impact assessment methodology. *International Data Privacy Law* 5, 1 (2015), 40–53.
[216] Nicholas D. Writer, Shazeda Ahmed, Natasha E. Bajema, Samuel Bendett, Benjamin A. Chang, Rogier Creemers, Chris C. Demchak, et al. 2019. *Artificial Intelligence, China, Russia, and the Global Order Technological, Political, Global, and Creative Perspectives*. Technical Report. Air University Press, Maxwell AFB.
[217] Xiaoyuan Xie, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. 2011. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software* 84, 4 (2011), 544–558.
[218] Kai Xu, Dae Hoon Park, Chang Yi, and Charles Sutton. 2018. Interpreting deep classifier by visual distillation of dark knowledge. *arXiv preprint arXiv:1803.04042* (2018).
[219] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. 2018. Global model interpretation via recursive partitioning. In *Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications, the IEEE 16th International Conference on Smart City, and the IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS'18)*. IEEE, Los Alamitos, CA, 1563–1570.
[220] Karen Yeung. 2017. 'Hypernudge': Big data as a mode of regulation by design. *Information, Communication & Society* 20, 1 (2017), 118–136.
[221] Heung Youl Youm. 2020. An overview of de-identification techniques and their standardization directions. *IEICE Transactions on Information and Systems* 103, 7 (2020), 1448–1461.
[222] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953* (2018). https://arxiv.org/abs/1812.02953.
[223] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 962–970.
[224] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2018. Linking artificial intelligence principles. *arXiv preprint arXiv:1812.04814* (2018). https://arxiv.org/abs/1812.04814.
[225] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[226] Yunfeng Zhang, Rachel Bellamy, and Kush Varshney. 2020. Joint optimization of AI fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 400–406.
[227] Guannan Zhao, Bo Zhou, Kaiwen Wang, Rui Jiang, and Min Xu. 2018. Respond-CAM: Analyzing deep models for 3D imaging data by visualizations. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. 485–492.
[228] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 119–134.

Received December 2020; revised September 2021; accepted October 2021