Trustworthy AI: From Principles to Practices

BO LI, JD Technology, China and Tsinghua University, China
PENG QI, Amazon AWS AI Labs, USA
BO LIU, Walmart Inc., USA
SHUAI DI, JD Technology, China
JINGEN LIU, JD Technology, USA
JIQUAN PEI, JD Technology, China
JINFENG YI, Frontis.AI, China
BOWEN ZHOU, Tsinghua University, China and Frontis.AI, China

The rapid development of Artificial Intelligence (AI) technology has enabled the deployment of various systems based on it. However, many current AI systems are found vulnerable to imperceptible attacks, biased against underrepresented groups, lacking in user privacy protection. These shortcomings degrade user experience and erode people's trust in all AI systems. In this review, we provide AI practitioners with a comprehensive guide for building trustworthy AI systems. We first introduce the theoretical framework of important aspects of AI trustworthiness, including robustness, generalization, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability. To unify currently available but fragmented approaches toward trustworthy AI, we organize them in a systematic approach that considers the entire lifecycle of AI systems, ranging from data acquisition to model development, to system development and deployment, finally to continuous monitoring and governance. In this framework, we offer concrete action items for practitioners and societal stakeholders (e.g., researchers, engineers, and regulators) to improve AI trustworthiness. Finally, we identify key opportunities and challenges for the future development of trustworthy AI systems, where we identify the need for a paradigm shift toward comprehensively trustworthy AI systems.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Trustworthy AI, robustness, generalization, explainability, transparency, reproducibility, fairness, privacy protection, accountability

ACM Reference format:
Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* 55, 9, Article 177 (January 2023), 46 pages.
https://doi.org/10.1145/3555803

Authors' addresses: B. Li, JD Technology, Beijing, China and Tsinghua University, Beijing, China; email: prclibo@gmail.com; P. Qi, Amazon AWS AI Labs, Seatle, WA, USA; email: qipeng.thu@gmail.com; B. Liu, Walmart Inc., Mountain View, CA, USA; email: kfliubo@gmail.com; S. Di, JD Technology, Beijing, China; email: dishuai@jd.com; J. Liu, JD Technology, Mountain View, CA, USA; email: jingenliu@gmail.com; J. Pei, JD Technology, Beijing, China; email: peijiquan@jd.com; J. Yi, Frontis.AI, Beijing, China; email: jinfengyi.ustc@gmail.com; B. Zhou (corresponding author), Tsinghua University, Beijing, China and Frontis.AI, Beijing, China; email: zhoubowen@tsinghua.edu.cn.

0360-0300/2023/01-ART177 $15.00
https://doi.org/10.1145/3555803

1 INTRODUCTION

The rapid development of **Artificial Intelligence (AI)** continues to provide significant economic and social benefits to society. With the widespread application of AI in areas such as transportation, finance, medicine, security, and entertainment, there is rising societal awareness that we need these systems to be trustworthy. This is because the breach of stakeholders' trust can lead to severe societal consequences given the pervasiveness of these AI systems. Such breaches can range from biased treatment by automated systems in hiring and loan decisions [49, 146] to the loss of human life [52]. By contrast, AI practitioners, including researchers, developers, and decision-makers, have traditionally considered system performance (i.e., accuracy) to be the main metric in their workflows. This metric is far from sufficient to reflect the trustworthiness of AI systems. Various aspects of AI systems beyond system performance should be considered to improve their trustworthiness, including but not limited to their robustness, algorithmic fairness, explainability, and transparency.

While most active academic research on AI trustworthiness has focused on the algorithmic properties of models, advancements in algorithmic research alone is insufficient for building trustworthy AI products. From an industrial perspective, the lifecycle of an AI product consists of multiple stages, including data preparation, algorithmic design, development, and deployment as well as operation, monitoring, and governance. Improving trustworthiness in any single aspect (e.g., robustness) involves efforts at multiple stages in this lifecycle, e.g., data sanitization, robust algorithms, anomaly monitoring, and risk auditing. On the contrary, the breach of trust in any single link or aspect can undermine the trustworthiness of the entire system. Therefore, AI trustworthiness should be established and assessed systematically throughout the lifecycle of an AI system.

In addition to taking a holistic view of the trustworthiness of AI systems over all stages of their lifecycle, it is important to understand the big picture of different aspects of AI trustworthiness. In addition to pursuing AI trustworthiness by establishing requirements for each specific aspect, we call attention to the combination of and interaction between these aspects, which are important and underexplored topics for trustworthy real-world AI systems. For instance, the need for data privacy might interfere with the desire to explain the system output in detail, and the pursuit of algorithmic fairness may be detrimental to the accuracy and robustness experienced by some groups [284, 361]. As a result, trivially combining systems to separately improve each aspect of trustworthiness does not guarantee a more trustworthy and effective end result. Instead, elaborated joint optimization and tradeoffs between multiple aspects of trustworthiness are necessary [47, 158, 331, 361, 380].

These facts suggest that a systematic approach is necessary to shift the current AI paradigm toward trustworthiness. This requires awareness and cooperation from multi-disciplinary stakeholders who work on different aspects of trustworthiness and different stages of the system's lifecycle. We have recently witnessed important developments in multi-disciplinary research on trustworthy AI. From the perspective of technology, trustworthy AI has promoted the development of adversarial learning, private learning, and the fairness and explainability of **machine learning (ML)**. Some recent studies have organized these developments from the perspective of either research [182, 218, 357] or engineering [57, 62, 199, 338, 353]. Developments in non-technical areas have also been reviewed in a few studies, and involve guidelines [145, 178, 294], standardization [210], and management processes [31, 274, 301]. We have conducted a detailed analysis of the various reviews, including algorithmic research, engineering practices, and institutionalization, in Section A.2 in the appendix. These fragmented reviews have mostly focused on specific views of trustworthy AI. To synchronize these diverse developments in a systematic view, we organize multi-disciplinary knowledge in an accessible manner for AI practitioners, and provide actionable and systematic guidance in the context of the lifecycle of an industrial system to build trustworthy AI systems. Our main contributions are as follows:

- We dissect the entire lifecycle of the development and deployment of AI systems in industrial applications, and discuss how AI trustworthiness can be enhanced at each stage—from data to AI models and from system deployment to its operation. We propose a systematic framework to organize the multi-disciplinary and fragmented approaches toward trustworthy AI and propose pursuing it as a continuous workflow to incorporate feedback at each stage of the lifecycle of the AI system.
- We dissect the entire development and deployment lifecycle of AI systems in industrial applications and discuss how AI trustworthiness can be enhanced at each stage—from data to AI models and from system deployment to its operation. We propose a systematic framework to organize the multi-disciplinary and fragmented approaches toward trustworthy AI and further propose to pursue AI trustworthiness as a continuous workflow to incorporate feedback at each stage of the AI system lifecycle. We also analyze the relationship between different aspects of trustworthiness in practice (mutual enhancement and, sometimes, trade-offs). The aim is to provide stakeholders of AI systems, such as researchers, developers, operators, and legal experts, with an accessible and comprehensive guide to quickly understand the approaches toward AI trustworthiness (Section 3).
- We discuss outstanding challenges facing trustworthy AI on which the research community and industrial practitioners should focus in the near future. We identify several key issues, including the need for a deeper and fundamental understanding of several aspects of AI trustworthiness (e.g., robustness, fairness, and explainability), the importance of user awareness, and the promotion of inter-disciplinary and international collaboration (Section 4).

With these contributions, we aim to provide the practitioners and stakeholders of AI systems not only with a comprehensive introduction to the foundations and future of AI trustworthiness but also with an operational guidebook for how to construct AI systems that are trustworthy.

2 AI TRUSTWORTHINESS: BEYOND PREDICTIVE ACCURACY

The success of ML technology in the past few decades has largely benefited from the accuracy-based performance measurements. By assessing task performance based on quantitative accuracy or loss, training AI models becomes tractable in the sense of optimization. Meanwhile, predictive accuracy is widely adopted to indicate the superiority of an AI product over others. However, with the recent widespread applications of AI, the limitation of an accuracy-only measurement has been exposed to a number of new challenges, ranging from malicious attacks against AI systems to misuses of AI that violate human values. To solve these problems, the AI community has realized in the last decade that factors beyond accuracy should be considered and improved when building an AI system. A number of enterprises [57, 62, 136, 166, 254, 338], academia [122, 199, 218, 301, 322], public sectors, and organizations [9, 210, 334] have recently identified these factors and summarized them as principles of AI trustworthiness. They include robustness, security, transparency, fairness, and safety [178]. Comprehensive statistics relating to and comparisons between these principles have been provided in References [145, 178]. In this article, we study the representative principles that have recently garnered wide interest and are closely related to practical applications. These principles can be categorized as follows:

- We consider representative requirements that pertain to technical challenges faced by current AI systems. We review aspects that have sparked wide interest in recent technical studies, including robustness, explainability, transparency,[1] reproducibility, and generalization.
- We consider ethical requirements with broad concerns in recent literature [9, 57, 121, 145, 178, 199, 218, 301, 334], including fairness, privacy, and accountability.

[1]We follow Reference [333] to exclude transparency as an ethical requirement.

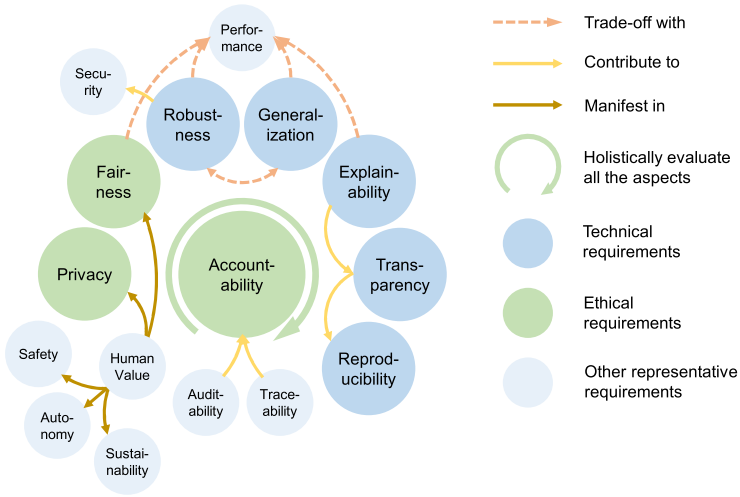

Fig. 1. The relation between different aspects of AI trustworthiness discussed in this survey. Note that implicit interaction widely exists between aspects, and we cover only representative explicit interactions.

In this section, we illustrate the motivation for and definition of each requirement. We also survey approaches to the evaluation of each requirement. It should also be noted that the selected requirements are not orthogonal, and some of them are closely correlated. We explain the relationships with the corresponding requirements in this section. We also use Figure 1 to visualize the relationships between aspects, including tradeoffs, contributions, and manifestation.

2.1 Robustness

In general, robustness refers to the ability of an algorithm or system to deal with execution errors, erroneous inputs, or unseen data. Robustness is an important factor affecting the performance of AI systems in empirical environments. The lack of robustness might also cause unintended or harmful behavior by the system, thus diminishing its safety and trustworthiness. In the context of ML systems, the term robustness is applicable to a diversity of situations. In this review, we non-exhaustively summarize the robustness of an AI system by categorizing vulnerabilities at the levels of data, algorithms, and systems, respectively.

Data. With the widespread application of AI systems, the environment in which an AI model is deployed becomes more complicated and diverse. If an AI model is trained without considering the diverse distributions of data in different scenarios, then its performance might be significantly affected. Robustness against distributional shifts has been a common problem in various applications of AI [19]. In high-stake applications, this problem is even more critical owing to its negative effect on safety and security. For example, in the field of autonomous driving, besides developing a perceptual system working in sunny scenes, academia and the industry are using numerous development and testing strategies to enhance the perceptual performance of the vehicles in nighttime/rainy scenes to guarantee the system's reliability under a variety of weather conditions [318, 382].

Algorithms. It is widely recognized that AI models might be vulnerable to attacks by adversaries with malicious intentions. Among the various forms of attacks, the adversarial attack and defenses against it have raised concerns in both academia and the industry in recent years. Literature has categorized the threat of adversarial attacks in several typical aspects and proposed various

defense approaches [12, 69, 213, 304, 373]. For example, in Reference [340], adversarial attacks were categorized with respect to the attack timing. *Decision-time attack* perturbs input samples to mislead the prediction of a given model so that adversary could evade security checks or impersonate victims. *Training-time attack* injects carefully designed samples into the training data to change the system's response to specific patterns and is also known as *poisoning attack*. Considering the practicality of attacks, it is also useful to note the differences of attacks in terms of the spaces in which they are carried out. Conventional studies have mainly focused on *feature space attacks*, which are generated directly as the input features of a model. In many practical scenarios, the adversaries can modify only the input entity to indirectly produce attack-related features. For example, it is easy for someone to wear adversarial pattern glasses to evade a face verification system but difficult to modify the image data in memory. Studies on producing realizable entity-based attacks (*problem space attacks*) have recently garnered increasing interest [325, 358]. Algorithm-level threats might exist in various forms, in addition to directly misleading AI models. *Model stealing* (a.k.a. the *exploratory attack*) attempts to steal knowledge about models. Although it does not directly change model behavior, the stolen knowledge has significant value for generating adversarial samples [329].

Systems. System-level robustness against illegal inputs should also be carefully considered in realistic AI products. The cases of illegal inputs can be extremely diverse in practical situations. For example, an image with a very high resolution might cause an imperfect image recognition system to hang. A lidar perception system for an autonomous vehicle might perceive laser beams emitted by lidars in other vehicles and produce corrupted inputs. The *presentation attack* [275] (a.k.a. *spoof attack*) is another example that has generated wide concerns in recent years. It fakes inputs by, for example, photos or masks to fool biometric systems.

Various approaches have been explored to prevent vulnerabilities in AI systems. The objective of defense can be either proactive or reactive [227]. A proactive defense attempts to optimize the AI system to be more robust to various inputs while a reactive defense aims at detecting potential security issues, such as changing distributions or adversarial samples. Representative approaches to improve the robustness of an AI system are introduced in Section 3.

Evaluation. Evaluating the robustness of an AI system serves as an important means of avoiding vulnerabilities and controlling risks. We briefly describe two groups of evaluations: robustness test and mathematical verification.

Robustness test. Testing has served as an essential approach to evaluate and enhance the robustness not only of conventional software but also of AI systems. Conventional functional test methodologies, such as *the monkey test* [115], provide effective approaches to evaluating the system-level robustness. Moreover, as will be introduced in Section 3.3.1, software testing methodologies have recently been extended to evaluate robustness against adversarial attacks [226, 260].

In comparison with functional test, performance test, i.e., benchmarking, are more widely adopted approach in the area of ML to evaluate system performance along various dimensions. Test datasets with various distributions are used to evaluate the robustness of data in ML research. In the context of adversarial attacks, the minimal adversarial perturbation is a core metric of robustness, and its empirical upper bound, a.k.a. empirical robustness, on a test dataset has been widely used [65, 312]. From the attacker's perspective, the rate of success of an attack also intuitively measures the robustness of the system [312].

Mathematical verification. Inherited from the theory of formal method, certified verification of the adversarial robustness of an AI model has led to growing interest. For example, adversarial robustness can be reflected by deriving a non-trivial and certified lower bound of the minimum distortion to an attack on an AI model [51, 379]. We introduce this direction in Section 3.2.1.

2.2 Generalization

Generalization has long been a source of concern in ML models. It represents the capability to distill knowledge from limited training data to make accurate predictions regarding unseen data [133]. Although generalization is not a frequently mentioned direction in the context of trustworthy AI, we find that its impact on AI trustworthiness should not be neglected and deserves specific discussion. On the one hand, generalization requires that AI systems make predictions on realistic data, even on domains or distributions on which they are not trained [133]. This significantly affects the reliability and risk of practical systems. On the other hand, AI models should be able to generalize without the need to exhaustively collect and annotate large amounts of data for various domains [343, 391], so that the deployment of AI systems in a wide range of applications is more affordable and sustainable.

In the field of ML, the canonical research on generalization theory has focused on the prediction of unseen data, which typically share the same distribution as the training data [133]. Although AI models can achieve a reasonable accuracy on training datasets, it is known that a gap (a.k.a generalization gap) exists between their training and testing accuracies. Approaches in different areas, ranging from statistic learning to deep learning, have been studied to analyze this problem and enhance the model generalization. Typical representatives like cross-validation, regularization, and data augmentation can be found in many ML textbooks [133].

The creation of a modern data-driven AI model requires a large amount of data and annotations in the training stage. This leads to a high cost for manufacturers and users for re-collecting and re-annotating data to train the model for each task. The cost highlights the need to generalize the knowledge of a model to different tasks, which not only reduces data cost but also improves model performance in many cases. Various directions of research have been explored to address knowledge generalization under different scenarios and configurations within the paradigm of transfer learning [255, 350]. We review the representative approaches in Section 3.2.2.

The inclusive concept of generalization is closely related to other aspects of AI trustworthiness, especially robustness. In the context of ML, the robustness against distributional shifts (Section 2.1) is also considered a problem of generalization. This implies that the requirements of robustness and generalization have some overlapping aspects. The relationship between adversarial robustness and generalization is more complicated. As demonstrated in Reference [362], an algorithm that is robust against small perturbations has better generalization. Recent research [271, 331], however, has noted that improving robustness by adversarial training may reduce the testing accuracy and leads to worse generalization. To explain this phenomenon, Reference [116] has argued that the adversarial robustness corresponds to different data distributions that may hurt a model's generalization.

Evaluation. Benchmarking on test datasets with various distributions is a widely used approach to evaluate the generalization of an AI model in realistic scenarios. A summary of commonly used datasets and benchmarks for domain generalization can be found in Reference [391] and covers the tasks of object recognition, action recognition, segmentation, and face recognition.

In terms of theoretical evaluation, past ML studies have developed rich approaches to measure the bounds of a model's generalization error. For example, Rademacher complexity [35] is commonly used to determine how well a model can fit a random assignment of class labels. In addition, the **Vapnik–Chervonenkis (VC)** dimension [337] is a measure of the capacity/complexity of a learnable function set. A larger number of VC dimensions indicates a higher capacity.

Advances in the DNN has led to new developments in the theory of generalization. Reference [377] observed that modern deep learning models can achieve a generalization gap despite their massive capacity. This phenomenon has led to academic discussions on the generalization of the **Deep Neural Network (DNN)** [23, 39]. For example, Reference [39] examined generalization

from the perspective of the bias–variance tradeoff to explain and evaluate the generalization of the DNN.

2.3 Explainability and Transparency

The opaqueness of complex AI systems has led to widespread concerns in academia, the industry, and society at large. The problem of how DNNs outperform other conventional ML approaches has been puzzling researchers [24]. From the perspective of practical systems, there is a demand among users for the right to know the intention, business model, and technological mechanism of AI products [9, 135]. A variety of studies have addressed these problems in terms of nomenclature including interpretability, explainability, and transparency [5, 24, 47, 141, 216, 250] and have delved into different definitions. To make our discussion more concise and targeted, we narrow the coverage of explainability and transparency to address the above concerns in theoretical research and practical systems, respectively.

- **Explainability** addresses to understand how an AI model makes decision [24].
- **Transparency** considers AI as a software system, and seeks to disclose information regarding its entire lifecycle (cf., "operate transparently" in Reference [9]).

2.3.1 Explainability. Explainability, i.e., understanding how an AI model makes its decision, stays at the core place of modern AI research and serves as a fundamental factor that determines the trust in AI technology. The motivation for the explainability of AI comes from various aspects [24, 25]. From the perspective of scientific research, it is meaningful to understand all intrinsic mechanisms of the data, parameters, procedures, and outcomes in an AI system. The mechanisms also fundamentally determine AI trustworthiness. From the perspective of building AI products, there exist various practical requirements on explainability. For operators like bank executives, explainability helps understand the AI credit system to prevent potential defects in it [25, 184]. Users like loan applicants are interested to know why they are rejected by the model, and what they can do to qualify [25]. See Reference [25] for a detailed analysis of the various motivations of explainability.

Explaining ML models has been an active topic not only in ML research but also in psychological research in the past five years [5, 24, 47, 141, 216, 250]. Although the definition of the explainability of an AI model is still an open question, research has sought to address this problem from the perspectives of AI [141, 285] and psychology [144, 245]. In summary, the relevant studies divide explainability into two levels to explain it.

- **Model explainability by design.** A series of fully or partially explainable ML models have been designed in the past half-century of ML research. Representatives include linear regression, trees, the **k-nearest neighbors (KNN)**, rule-based learners, generalized additive model, and Bayesian models [24]. The design of explainable models remains an active area in ML.
- **Post hoc model explainability.** Despite the good explainability of the above conventional models, more complex models such as the DNN or **Gradient Boosting Decision Tree (GDBT)** have exhibited better performance in recent industrial AI systems. Because the relevant approaches still cannot holistically explain these complex models, researchers have turned to post hoc explanation. It addresses a model's behavior by analyzing its input, intermediate result, and output. A representative category in this vein approximates the decision surface either globally or locally by using an explainable ML model, i.e., *explainer*, such as a linear model [225, 279] and rules [140, 280]. For deep learning models like the **Convolutional Neural Network (CNN)** or the transformer, the inspection of intermediate features is a widely used means of explaining model behavior [332, 366].

Approaches to explainability are an active area of work in ML and have been comprehensively surveyed in a variety of studies [24, 47, 141, 250]. Representative algorithms to achieve the above two levels of explainability are reviewed in Section 3.2.3.

Evaluation. In addition to the problem of explaining AI models, a unified evaluation of explainability has been recognized as a challenge. A major reason for this lies in the ambiguity of the psychological outlining of explainability. To sidestep this problem, a variety of studies have used qualitative metrics to evaluate explainability with human participation. Representative approaches include the following:

- **Subjective human evaluation.** The methods of evaluation in this context include interviews, self-reports, questionnaires, and case studies that measure, e.g., user satisfaction, mental models, and trust [144, 155, 267].
- **Human–AI task performance.** In tasks performed with human–AI collaboration, the collaborative performance is significantly affected by the human understanding of the AI collaborator and can be viewed as a reflection of the quality of explanation [249]. This evaluation has been used for the development of, for instance, recommendation systems [198] and data analysis [132].

In addition, if explainability can be achieved by an explainer, then the performance of the latter, such as in terms of the precision of approximation (*fidelity* [140, 279, 280]), can be used to indirectly and quantitatively evaluate explainability [16].

Despite the above evaluations, a direct quantitative measurement of explainability remains a problem. Some naive measurements of model complexity, like tree depth [46] and the size of the rule set [202], have been studied as surrogate explainability metrics in previous work. We believe that a unified quantitative metric lies at the very heart of fundamental AI research. Recent research on the complexity of ML models [162] and their cognitive functional complexity [347] may inspire future research on a unified quantitative evaluation metric.

2.3.2 Transparency. Transparency requires the disclosure of the information on a system, and has long been a recognized requirement in software engineering [89, 207]. In the AI industry, this requirement naturally covers the lifecycle of an AI system and helps stakeholders confirm that appropriate design principles are reflected in it. Consider a biometric system for identification as an example. Users are generally concerned with the purpose for which their biometric information is collected and how it is used. Business operators are concerned with the accuracy and robustness against attacks so that they can control risks. Government sectors are concerned with whether the AI system follows guidelines and regulations. On the whole, transparency serves as a basic requirement to build the public's trust in AI systems [22, 178, 189].

To render the lifecycle of an AI system transparent, a variety of information regarding its creation needs to be disclosed, including the design purposes, data sources, hardware requirements, configurations, working conditions, expected usage, and system performance. A series of studies have examined disclosing this information through appropriate documentation [22, 129, 156, 246, 265]. This is discussed in Section 3.5.1. The recent trend of open source systems also significantly contributes to the algorithmic transparency of AI systems.

Owing to the complex and dynamic internal procedure of an AI system, facts regarding its creation are not sufficient to fully reveal its mechanism. Hence, the transparency of the runtime process and decision-making should also be considered in various scenarios. For an interactive AI system, an appropriately designed user interface serves as an important means to disclose the underlying decision procedure [10]. In many safety-critical systems, such as autonomous driving vehicles, logging systems [29, 261, 369] are widely adopted to trace and analyze the system execution.

Evaluation. Although a unified quantitative evaluation is not yet available, the qualitative evaluation of transparency has undergone recent advances in the AI industry. Assessment checklists [10, 292] have been regarded as an effective means to evaluate and enhance the transparency of a system. In the context of the psychology of users or the public, user studies or A/B test can provide a useful evaluation based on user satisfaction [249].

Quality evaluations of AI documentation have also been explored in recent years. Some studies [22, 129, 156, 246, 273] have proposed standard practices to guide and evaluate the documentation of an AI system. Reference [265] summarized the general qualitative dimensions for more specified evaluations.

2.4 Reproducibility

Modern AI research involves both mathematical derivation and computational experiments. The reproducibility of these computational procedures serves as an essential step to verify AI research. In terms of AI trustworthiness, this verification facilitates the detection, analysis, and mitigation of potential risks in an AI system, such as a vulnerability on specific inputs or unintended bias. With the gradual establishment of the open cooperative ecosystem in the AI research community, reproducibility is emerging as a concern among researchers and developers. In addition to enabling the effective verification of research, reproducibility allows the community to quickly convert the latest approaches into practice or conduct follow-up research.

There has been a new trend in the AI research community to regard reproducibility as a requirement when publicizing research [142]. We have seen major conferences, such as **Neural Information Processing Systems (NeurIPS)**, **International Conference on Machine Learning (ICML)**, and **ACM Multimedia (ACMMM)**, introduce reproducibility-related policies or programs [263] to encourage the reproducibility of works. To obtain a clear assessment, degrees of reproducibility have been studied in such works as the ACM Artifact Review and References [106, 143]. For example, in Reference [143], the lowest degree of reproducibility requires the exact replication of an experiment with the same implementation and data, while a higher degree requires using different implementations or data. Beyond the basic verification of research, a higher degree of reproducibility promotes a better understanding of the research by distinguishing among the key factors influencing effectiveness.

Some recently developed large scale pre-trained AI models, such as **Generative Pre-trained Transformer 3 (GPT-3)** and **Bidirectional Encoder Representations from Transformers (BERT)**, are representative of the challenges to the reproducibility of AI research. The creation of these models involves specifically designed data collection strategies, efficient storage of big data, communication and scheduling between distributed clusters, algorithm implementation, appropriate software and hardware environments, and other kinds of knowhow. The reproducibility of such a model should be considered over its entire lifecycle. In recent studies on the reproducibility of ML, this requirement has been decomposed into the reproducibility of data, methods, and experiments [142, 143, 169], where the latter range over a series of lifecycle artifacts such as code, documentation, software, hardware, and deployment configuration. Based on this methodology, an increasing number of ML platforms are being developed that assist researchers and developers better track the lifecycle in a reproducible manner [169, 374].

Evaluation. Reproducibility checklists have been recently widely adopted in ML conferences to assess the reproducibility of submissions [263]. Beyond the replication of experiments in publication, References [142, 143] also specified checklists to evaluate reproducibility at varying degrees. In addition to checklists, mechanisms such as challenges to reproducibility and paper tracks of reproducibility have been adopted to evaluate the reproducibility of publications [118, 263]. To

quantitatively evaluate reproducibility in the context of challenges, a series of quantitative metrics have been studied. For example, References [53, 118] designed metrics to quantify how closely an information retrieval system can be reproduced to its origins.

2.5 Fairness

When AI systems help us in areas such as hiring, financial risk assessment, and face identification, systematic unfairness in their decisions might have negative social ramifications (e.g., underprivileged groups might experience systematic disadvantage in hiring decisions [49], or be disproportionally impacted in criminal risk profiling [104, 146, 161]). This not only damages the trust that various stakeholders have in AI but also hampers the development and application of AI technology for the greater good. Therefore, it is important that practitioners keep in mind the fairness of AI systems to avoid instilling or exacerbating social bias [66, 105, 242].

A common objective of fairness in AI systems is to mitigate the effects of biases. The mitigation is non-trivial, because the biases can take various forms, such as data bias, model bias, and procedural bias, in the process of developing and applying AI systems [242]. Bias often manifests in the form of unfair *treatment* of different *groups* of people based on their protected information (e.g., gender, race, and ethnicity). Therefore, group identity (sometimes also called *sensitive variables*) and system response (prediction) are two factors influencing bias. Some cases also involve objective *ground truths* of a given task that one should consider when evaluating system fairness, e.g., whether a person's speech is correctly recognized or their face correctly identified.

Fairness can be applicable at multiple granularities of system behavior [66, 242, 339]. At each granularity, we might be concerned with *distributive fairness* or fairness of the outcome, or *procedural fairness* or fairness of the process (we refer the reader to Reference [137] for a more detailed discussion). In each case, we are commonly concerned with the aggregated behavior and the bias therein of an AI system, which is referred to as *statistical fairness* or *group fairness*. In certain applications, it is also helpful to consider *individual fairness* or *counterfactual fairness*, especially when the sensitive variable can be more easily decoupled from the other features that should justifiably determine the system's prediction [242]. While the former is more widely applicable to various ML tasks, e.g., speech recognition and face identification, the latter can be critical in cases like resume reviewing for candidate screening [44].

At the group level, researchers have identified three abstract principles to categorize different types of fairness [66]. We illustrate them with a simple example of hiring applicants from a population consisting of 50% male and 50% female applicants, where gender is the sensitive variable (examples adapted from References [339, 388]):

- **Independence.** This requires for the system outcome to be statistically independent of sensitive variables. In our example, this requires that the rate of admission of male and female candidates be equal (known as *demographic parity* [376]; see also *disparate impact* [117]).
- **Separation.** Independence does not account for a justifiable correlation between the ground truth and the sensitive variable (e.g., fewer female candidates might be able to lift 100-lb goods more easily than male candidates). Separation therefore requires that the independence principle hold, conditioned on the underlying ground truth. That is, if the job requires strength qualifications, then the rate of admission for qualified male and female candidates should be equal (known as *equal opportunity* [147]; see also *equal odds* [43] and *accuracy equity* [95]).
- **Sufficiency.** Sufficiency similarly considers the ground truth but requires that the true outcome and the sensitive variable be independent when conditioned on the same system prediction. That is, given the same hiring decision predicted by the model, we want the same ratio

of qualified candidates among male and female candidates (known as *test fairness* [80, 147]). This is closely related to model calibration [266].

Note that these principles are mutually exclusive under certain circumstances (e.g., independence and separation cannot both hold when the sensitive variable is correlated with the ground truth). Reference [187] has discussed the tradeoff between various fairness metrics. Furthermore, Reference [84] advocated an extended view of these principles, where the utility of the predicted and true outcomes is factored into consideration (e.g., the risk and cost of recidivism in violent crimes compared with the cost of detention), and can be correlated with the sensitive variables. We refer the reader to this work for a more detailed discussion.

Evaluation. Despite the simplicity of the abstract criteria outlined in the previous section, fairness can manifest in many different forms following these principles (see Reference [66, 356] for comprehensive surveys, and Reference [228] for AI ethics' checklists). We categorize metrics of fairness according to the properties of models and tasks to help the reader choose appropriate ones for their application:

Discrete vs. continuous variables. The task output, model prediction, and sensitive variables can all be discrete (e.g., classification and nationality), ranked (e.g., search engines, recommendation systems), or continuous (e.g., regression, classifier scores, age, etc.) in nature. An empirical correlation of discrete variables can be evaluated with standard statistical tools, such as correlation coefficients (Pearson/Kendall/Spearman) and **Analysis of Variance (ANOVA)**, while continuous variables often further require binning, quantization, or loss functions to evaluate fairness [66].

Loss function. The criteria of fairness often cannot be exactly satisfied given the limitations of empirical data (e.g., demographic parity between groups when hiring only three candidates). Loss functions are useful in this case to gauge how far we are from empirical fairness. The choice of loss function can be informed by the nature of the variables of concern: If the variables represent probabilities, then likelihood ratios are more meaningful (e.g., disparate impact [117]); for real-valued regression, the difference between mean distances to the true value aggregated over each group might be used instead to indicate whether we model one group significantly better than another [59].

Multiple sensitive variables. In many applications, the desirable AI system should be fair to more than one sensitive variables (e.g., the prediction of risk posed by a loan should be fair in terms of both gender and ethnicity; *inter alia*, a recommendation system should ideally be fair to both users and recommendees). One can either form a tradeoff of "marginal fairness" between these variables when they are considered one at a time, i.e., evaluate the fairness of each variable separately and combine the loss functions for final evaluation, or explore the full Cartesian product [307] of all variables to achieve joint fairness, which typically requires more empirical observations but tends to satisfy stronger ethical requirements.

2.6 Privacy Protection

Privacy protection mainly refers to protecting against unauthorized use of the data that can directly or indirectly identify a person or household. These data cover a wide range of information, including name, age, gender, face image, fingerprints, and so on. Commitment to privacy protection is regarded as an important factor determining the trustworthiness of an AI system. The recently released AI ethics guidelines also highlight privacy as one of the key concerns [9, 178]. Government agencies are formulating a growing number of policies to regulate the privacy of data. The **General Data Protection Regulation (GDPR)** is a representative legal framework, which pushes enterprises to take effective measures for user privacy protection.

In addition to internal privacy protection within an enterprise, recent developments in data exchange across AI stakeholders has yielded new challenges for privacy protection. For example, when training a medical AI model, each healthcare institution typically only has data from the local residents, which might be insufficient. This leads to the demand to collaborate with other institutions and jointly train a model [299] without leaking private information across institutions.

Existing protection techniques penetrate the entire lifecycle of AI systems to address rising concerns about privacy. In Section 3, we briefly review techniques to protect the privacy in data collection and processing, model training (Section 3.2.5), and model deployment (Section 3.4.4). The realization of privacy protection is also related to other aspects of trustworthy AI. For example, the transparency principle is widely used in AI systems. It informs users of personal data collection and enables privacy settings. In the development of privacy-preserving ML software, such as federated learning (e.g., FATE and PySyft), open source is a common practice to increase transparency and certify the protectiveness of the system.

Evaluation. Laws for data privacy protection like the GDPR require **data protection impact assessment (DPIA)** if any data processing poses a risk to data privacy. Measures have to be taken to address the risk-related concerns and demonstrate compliance with the law [10]. Data privacy protection professionals and other stakeholders need to be involved to evaluate it.

Previous research has devised various mathematical methods to formally verify the protectiveness of privacy-preserving approaches. Typical verification can be conducted under assumptions such as semi-honest security, which implies all participating parties follow a protocol to perform the computational task but may try to infer the data of other parties from the intermediate results of computation (e.g., Reference [215]). A stricter assumption is the malicious attack assumption, where each participating party need not follow the given protocol, and can take any possible measure to infer the data [214].

In practical scenarios, the empirical evaluation of the risk of leakage of privacy is usually considered [283, 360]. For example, Reference [283] showed that 15 demographic attributes were sufficient to render 99% of the participants unique. An assessment of such data re-identification intuitively reflects protectiveness when designing a data collection plan.

2.7 Accountability: A Holistic Evaluation throughout the above Requirements

We have described a series of requirements to build trustworthy AI. Accountability addresses the regulation on AI systems to follow these requirements. With gradually improving legal and institutional norms on AI governance, accountability becomes a crucial factor for AI to sustainably benefit society with trustworthiness [100].

Accountability runs through the entire lifecycle of an AI system and requires that the stakeholders of an AI system be obligated to justify their design, implementation, and operation as aligned with human values. At the level of executive, this justification is realized by considerate product design, reliable technique architecture, a responsible assessment of the potential impacts, and the disclosure of information on these aspects [209]. Note that in terms of information disclosure, transparency contributes the basic mechanism used to facilitate the accountability of an AI system [94, 100].

From accountability is also derived the concept of auditability, which requires the justification of a system to be reviewed, assessed, and audited [209]. Algorithmic auditing is a recognized approach to ensure the accountability of an AI system and assess its impact on multiple dimensions of human values [272]. See also Section 3.5.2.

Evaluation. Checklist-based assessments have been studied to qualitatively evaluate accountability and auditability [10, 315]. As mentioned in this section, we consider accountability to be

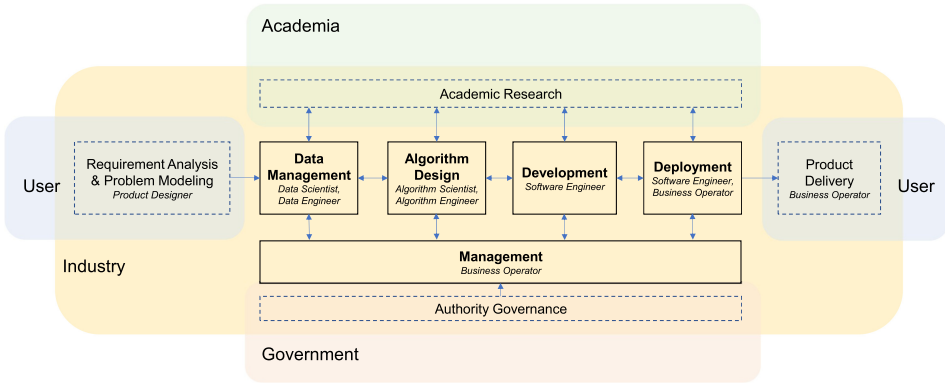

Fig. 2. The AI industry holds a connecting position to organize multi-disciplinary practitioners, including users, academia, and government in the establishment of trustworthy AI. In Section 3, we discuss the current approaches toward trustworthy AI in the five main stages of the lifecycle of an AI system, i.e., data preparation, algorithm design, development, deployment, and management.

the comprehensive justification of each concrete requirement of trustworthy AI. Its realization is composed of evaluations of these requirements over the lifecycle of an AI system [272]. Hence, the evaluation of accountability is reflected by the extent to which these requirements of trustworthiness and their impact can be evaluated.

3 TRUSTWORTHY AI: A SYSTEMATIC APPROACH

We have introduced the concepts relevant to trustworthy AI in Section 2. Since the early 2010s, diverse AI stakeholders have made efforts to enhance AI trustworthiness. In Section A in our appendix, we briefly review their recent practices in multi-disciplinary areas, including research, engineering, and regulation, and studies that are exemplars of industrial applications, including on face recognition, autonomous driving, and **Natural Language Processing (NLP)**. These practices have made important progress in improving AI trustworthiness. However, we find that this work remains insufficient from the industrial perspective. As depicted in Section 1 and Figure 2, the AI industry holds a position to connect multi-disciplinary fields for the establishment of trustworthy AI. This position requires that industrial stakeholders learn and organize these multi-disciplinary approaches and ensure trustworthiness throughout the lifecycle of AI.

In this section, we provide a brief survey of techniques used for building trustworthy AI products and organize them across the lifecycle of product development from an industrial perspective. As shown by the solid-lined boxes in Figure 2, the lifecycle of the development of a typical AI product can be partitioned into data preparation, algorithm design, development–deployment, and management [26]. We review several critical algorithms, guidelines, and government regulations that are closely relevant to the trustworthiness of AI products in each stage of their lifecycle, with the aim of providing a systematic approach and an easy-to-follow guide to practitioners from varying backgrounds to establish trustworthy AI. The approaches and literature mentioned in this section are summarized in Figure 3 and Table 1.

3.1 Data Preparation

Current AI technology is largely data driven. The appropriate management and exploitation of data not only improves the performance of an AI system but also affects its trustworthiness. In

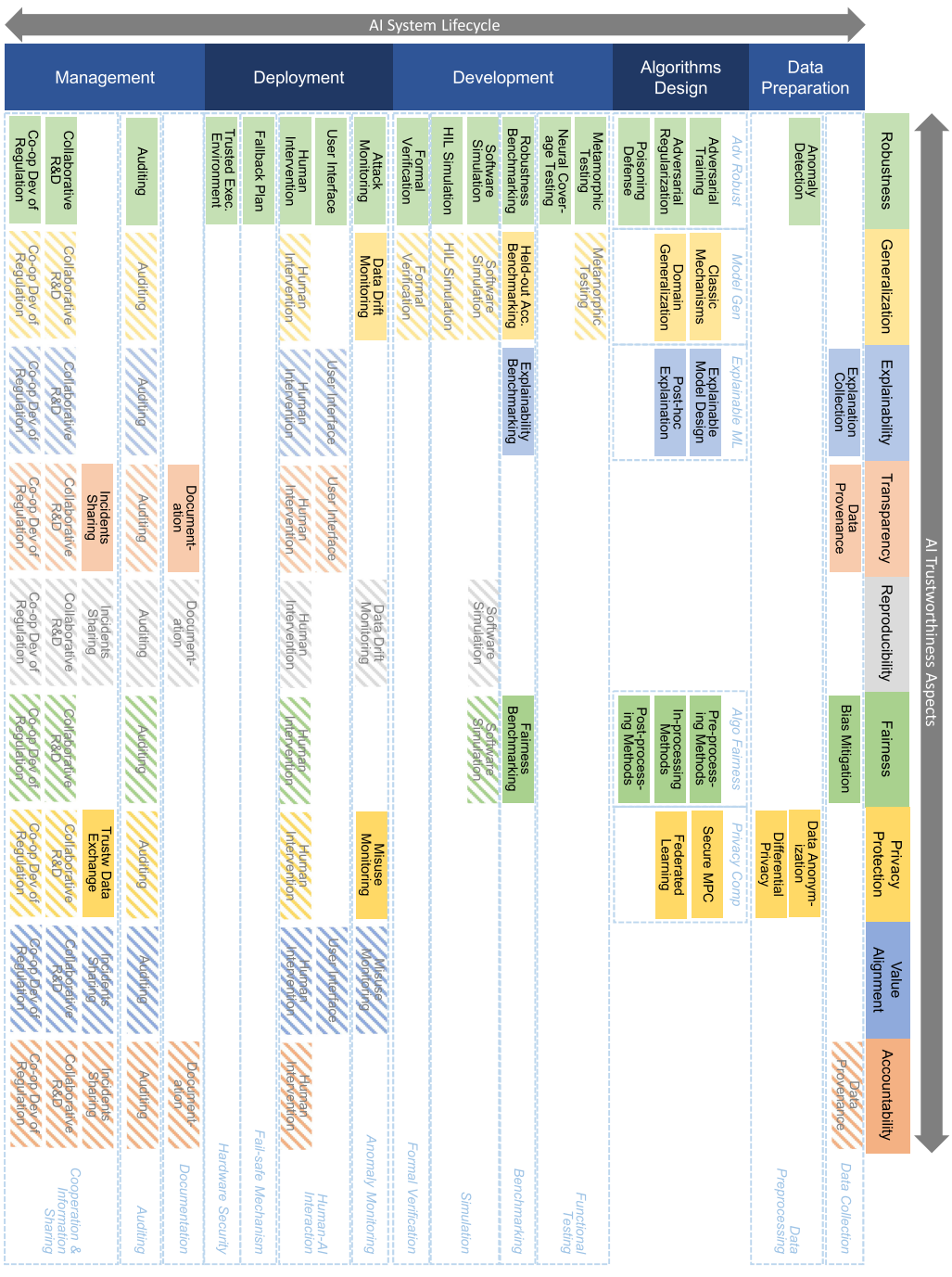

Fig. 3. A look-up table that organizes surveyed approaches to AI trustworthiness from different perspectives and in different stages of the lifecycle of the AI system. Some approaches can be used to improve AI trustworthiness from multiple aspects, and are in multiple columns. We dim these duplicated blocks here through stride-filling for better visualization. See the corresponding paragraphs in Section 3 for the details of the approaches.

Table 1. Representative Papers of Approaches or Research Directions Mentioned in Section 3

Lifecycle	Approaches	Literature
Data Preparation	Data Collection	Bias Mitigation [66, 242, 339]
		Data Provenance [154, 172]
	Data Preprocessing	Anomaly Detection [70, 81, 257, 316]
		Data Anonymozation [14, 37, 232, 232]
		Differential Privacy [107, 110, 177]
Algorithm Design	Adversarial Robustness	Adversarial Robustness [12, 19, 45, 69, 304, 346, 373]
		Poisoning defense [213]
	Explainability ML	Explainability ML [24, 47, 101, 101, 141, 250]
	Model Generalization	Model Generalization [1, 23, 38, 124, 133, 183, 377]
		Domain Generalization [343, 391]
	Algorithmic Fairness	Fairness and bias mitigation [66, 242, 339]
	Privacy Computing	SMPC [114, 387]
		Federated Learning [180, 365]
Development	Functional Testing	[164, 235, 381]
	Performance Benchmarking	[40, 56, 88, 93, 123, 127, 153, 238, 248, 307, 327]
	Simulation	[42, 50, 54, 103, 192, 205, 323, 324, 359]
	Formal Verification	[164, 298, 335]
Deployment	Anomaly Monitoring	[70, 257, 394]
	Human-AI Interaction	[72, 179, 301, 326]
	Fail-Safe Mechanism	[126, 160, 230, 252]
	Hardware Security	[264, 287, 364]
Management	Documentation	[11, 22, 41, 129, 156, 246]
	Auditing	[58, 228, 272, 277, 290, 292, 356]
	Cooperation	[27, 90]
Workflow	MLOps	[165, 233, 303]

For research directions that are widely studied, we provide the corresponding surveys for readers to refer to. For approaches or research directions without surveys available, we provide representative technical papers.

this section, we consider two major aspects of data preparation, i.e., data collection and data preprocessing. We also discuss the corresponding requirements of trustworthy AI.

3.1.1 Data Collection. Data collection is a fundamental stage of the lifecycle of AI systems. An elaborately designed data collection strategy can conduce to the enhancement of AI trustworthiness, such as in terms of fairness and explainability.

Bias mitigation. Training and evaluation data are recognized as a common source of bias for AI systems. Many types of bias might exist and plague fairness in data collection, requiring different processes and techniques to combat it (see Reference [242] for a comprehensive survey). Bias mitigation techniques during data collection can be divided into two broad categories: debias sampling and debias annotation. The former concerns the identification of data points to use or annotate, while the latter focuses on choosing the appropriate annotators.

When sampling data points to annotate, we note that a dataset reflecting the user population does not guarantee fairness, because statistical methods and metrics might favor majority groups. This bias can be further amplified if the majority group is more homogeneous for the task (e.g., recognizing speech in less-spoken accents can be naturally harder due to data scarcity [191]). System developers should therefore take task difficulty into consideration when developing and evaluating

fair AI systems. However, choosing the appropriate annotators is especially crucial for underrepresented data (e.g., when annotating speech recognition data, most humans are also poor at recognizing rarely heard accents). Therefore, care must be taken in selecting the correct experts, especially when annotating data from underrepresented groups, to prevent human bias from creeping into the annotated data.

Explanation collection. Aside from model design and development, data collection is also integral to building explainable AI systems. As will be mentioned in Section 3.2.3, adding an explanation task to the AI model can help explain the intermediate features of the model. This strategy is used in tasks like NLP-based reading comprehension by generating supporting sentences [332, 366]. To train the explanation task, it is helpful to consider collecting explanations or information that may not directly be part of the end task, either directly from annotators [354] or with the help of automated approaches [185].

Data Provenance. Data provenance requires recording the data lineage, including the source, dependencies, contexts, and steps of processing [306]. By tracking the data lineage at the highest resolution, data provenance can enhance the transparency, reproducibility, and accountability of an AI system [154, 172]. Moreover, recent research has shown that data provenance can be used to mitigate data poisoning [33], thus enhancing the robustness and security of an AI system. The technical realization of data provenance has been provided in Reference [154]. Tool chains [293] and documentation [129] guides have also been studied for specific scenarios involving AI systems. To ensure that the provenance is tamper proof, the blockchain has been recently considered as a promising tool to certify data provenance in AI [15, 96].

3.1.2 Data Preprocessing. Before feeding data into an AI model, data preprocessing helps remove inconsistent pollution of the data that might harm model behavior and sensitive information that might compromise user privacy.

Anomaly Detection. Anomaly detection (a.k.a. *outlier detection*) has long been an active area of ML [70, 81, 257, 316]. Due to the sensitivity of ML models to outlier data, data cleaning by anomaly detection serves as an effective approach to enhance performance. In recent studies, anomaly detection has been shown to be useful in addressing some requirements of AI trustworthiness. For example, fraudulent data can challenge the robustness and security of systems in areas such as banking and insurance. Various approaches have been proposed to address this issue by using anomaly detection [70]. The detection and mitigation of adversarial inputs is also considered to be a means to defend against evasion attacks and data poisoning attacks [12, 213, 304]. It is noteworthy that the effectiveness of detection in high dimensions (e.g., images) is still limited [64]. The mitigation of adversarial attacks is also referred to as *data sanitization* [71, 87, 258].

Data Anonymization (DA). DA modifies the data so that the protected private information cannot be recovered. Different principles of quantitative data anonymization have been developed, such as k-anonymity [288], (c, k)-safety [236], and δ-presence [253]. Data format-specific DA approaches have been studied for decades [171, 372, 386]. For example, private information in the form of graph data for social networks can be potentially contained in properties of the vertices of the graph, its link relationships, weights, or other graph metrics [390]. Ways of anonymizing such data have been considered in the literature [37, 220]. Specific DA methods have also been designed for relational data [262], set-valued data [151, 320], and image data [97, 239]. Guidelines and standards have been formulated for data anonymization, such as the US HIPAA and the UK ISB1523. Data pseudonymization [251] is also a relevant technique promoted by the GDPR. It replaces private information with non-identifying references.

Desirable data anonymization is expected to be immune from data de-anonymization or re-identification attacks that try to recover private information from anonymized data [111, 175].

For example, Reference [176] introduces several approaches into de-anonymize user information from graph data. To mitigate the risk of privacy leakage, an open source platform was provided in Reference [174] to evaluate the privacy protection-related performance of graph data anonymization algorithms against de-anonymization attacks.

Differential privacy (DP). DP shares information of groups within datasets while withholding individual samples [108–110]. Typical DP can be formally defined by ϵ-differential privacy. It measures the extent to which a (randomized) statistical function on the dataset reflects whether an element has been removed [108]. DP has been explored in various data publishing tasks, such as log data [159, 385], set-valued data [76], correlated network data [75], and crowdsourced data [278, 344]. It has also been applied to single- and multi-machine computing environments, and integrated with ML models for protecting model privacy [2, 120, 349]. Enterprises like Apple have used DP to transform user data into a form from which the true data cannot be reproduced [21]. In Reference [113], researchers proposed the RAPPOR algorithm that satisfies the definition of DP. The algorithm is used for crowdsourcing statistical analyses of user software. DP is also used to improve the robustness of AI models against adversarial samples [204].

3.2 Algorithm Design

A number of aspects of trustworthy AI have been addressed as algorithmic problems in the context of AI research, and have attracted widespread interest. We organize recent technical approaches by their corresponding aspects of AI trustworthiness, including robustness, explainability, fairness, generalization, and privacy protection, to provide a quick reference for practitioners.

3.2.1 Adversarial Robustness. The robustness of an AI model is significantly affected by the training data and the algorithms used. We describe several representative directions in this section. Comprehensive surveys can be found in the literature such as References [12, 19, 45, 69, 213, 304, 373].

Adversarial training. Since the discovery of adversarial attacks, it has been recognized that augmenting the training data with adversarial samples provides an intuitive approach for defense against them. This is typically referred to as adversarial training [134, 211, 346]. The augmentation can be carried out in a brute-force manner by feeding both the original data and adversarial samples during training [201], or by using a regularization term to implicitly represent adversarial samples [134]. Conventional adversarial training augments data with respect to specific attacks. It can defend against the corresponding attack but is vulnerable to other kinds of attacks. Various improvements have been studied to improve this defense [45, 229, 304]. Reference [328] augmented the training data with adversarial perturbations transferred from other models. It is shown to further provide defense against black-box attacks that do not require knowledge of the model parameters This can help defend against black-box attacks, which do not require knowledge of the model parameters. Reference [231] combined multiple types of perturbations into adversarial training to improve the robustness of the model against multiple types of attacks.

Adversarial regularization. In addition to the regularization term that implicitly represents adversarial samples, recent research further explores network structures or regularization to overcome the vulnerabilities of the DNN to adversarial attacks. An intuitive motivation for this regularization is to prevent the outcome of the network from changing dramatically in case of small input perturbations. For example, Reference [139] penalized large partial derivatives of each layer to improve the stability of its output. A similar regularization on gradients was adopted by Reference [286]. Parseval networks [82] train the network by imposing a regularization on the Lipschitz constant in each layer.

Certified robustness. Adversarial training and regularization improve the robustness of AI models in practice but cannot theoretically guarantee that the models work reliably. This problem has

prompted research to formally verify the robustness of models (a.k.a. *certified robustness*). Recent research on certified robustness has focused on robust training to deal with input perturbations. For example, CNN-Cert [51], CROWN [379], Fast-lin, and Fast-lip [352] aim to minimize an upper bound of the worst-case loss under given input perturbations. Reference [152] instead derives a lower bound on the input manipulation required to change the decision of the classifier, and uses it as a regularization term for robust training. To address the issue that the exact computation of such bounds is intractable for large networks, various relaxations or approximations, such as References [352, 378], have been proposed as alternatives to regularization. Note that the above research mainly optimizes robustness only locally near the given training data. To achieve certified robustness on unseen inputs as well, *global robustness* has recently attracted the interest of the AI community [77, 206].

It is also worth noting the recent trend of the intersection of certified robustness and the perspective of formal verification, which aims at developing rigorous mathematical specification and techniques of verification for assurances of software correctness [83]. A recent survey by Reference [335] provided a thorough review of the formal verification of neural networks.

Poisoning defense. Typical poisoning or backdoor attacks contaminate the training data to mislead model behavior. Besides avoiding suspicious data in the data sanitization stage, defensive algorithms against poisoning data are an active area [213]. The defense has been studied at different stages of a DNN model. For example, based on the observation that backdoor-related neurons are usually inactivated for benign samples, Reference [219] proposed pruning these neurons from a network to remove the hidden backdoor. Neural Cleanse [342] proactively discovers backdoor patterns in a model. The backdoor can then be avoided by the early detection of backdoor patterns from the data or retraining the model to mitigate the backdoor. The detection of backdoor attacks can also be carried out by analyzing the prediction on the model on specifically designed benchmarking inputs [194].

3.2.2 Model Generalization. Techniques of model generalization not only aim to improve model performance but also explore training AI models with limited data and at limited cost. We review representative approaches to model generalization, categorized as classic generalization and domain generalization.

Classic generalization mechanisms. As a fundamental principle of model generalization theory, the bias–variance tradeoff indicates that a generalized model should maintain a balance between underfitting and overfitting [39, 124]. For an overfitted model, reducing complexity/capacity might lead to better generalization. Consider the neural network as an example. Adding a bottleneck layer, which has fewer neurons than the layers both below and above, to it can help reduce model complexity and reduce overfitting.

Other than adjusting the architecture of the model, one can mitigate overfitting to obtain better generalization via various explicit or implicit regularizations, such as early stopping [370], batch normalization [167], dropout [309], data augmentation, and weight decay [196]. These regularizations are standard techniques to improve model generalization when the training data are much smaller in size than the number of model parameters [337]. They aim to push learning to a subspace of a hypothesis with manageable complexity and reduce model complexity [377]. However, [377] also observed that explicit regularization may improve generalization performance but is insufficient to reduce generalization error. The generalization of the deep neural network is thus still an open problem.

Domain generalization. A challenge for modern DNNs is their generalization of out-of-distribution data. This challenge arises from various practical AI tasks [343, 391] in the area of transfer learning [255, 350]. Domain adaptation [343, 391] aims to find domain-invariant features such that an algorithm can achieve similar performances across domains. As another example, the goal

of few-shots learning is to generalize model to new tasks using only a few examples [78, 348, 371]. Meta-learning [336] attempts to learn prior knowledge of generalization from a number of similar tasks. Feature similarity [190, 308] has been used as a representative type of knowledge prior in works such as the **Model-Agnostic Meta-Learning (MAML)** [119], reinforcement learning [212], and memory-augmented neural network [38, 291].

Model pre-training is a popular mechanism to leverage knowledge learned from other domains, and has recently achieved growing success in both academia and the industry. For example, in computer vision, an established successful paradigm involves pre-training models on a large-scale dataset, such as ImageNet, and then fine-tuning them on target tasks with fewer training data [131, 224, 375]. This is because pre-trained feature representation can be used to transfer information to the target tasks [375]. Unsupervised pre-training has recently been very successful in language processing (e.g., BERT [92] and GPT [269]) and computer vision tasks (e.g., **Momentum Contrast (MoCo)** [150] and **Sequence Contrastive learning (SeCo)** [368]). In addition, self-supervised learning provides a good mechanism to learn a cross-modal feature representation. These include the vision and language models VL-BERT [313] and Auto-CapTIONs [256]. To explain the effectiveness of unsupervised pre-training, [112] conducted a series of experiments to illustrate that it can drive learning to the basins of minima that yield better generalization.

3.2.3 Explainable ML. In this section, we review representative approaches for the two aspects of ML explainability mentioned in Section 2.3.1 and their application to different tasks.

Explainable ML model design. In spite of being recognized as disadvantageous in terms of performance, explainable models have been actively researched in recent years, and a variety of fully or partially explainable ML models have been studied to push their performance limit.

Self-explainable ML models. A number of self-explainable models have been studied in ML over the years. The representative ones include the KNN, linear/logistic regression, decision trees/rules, and probabilistic graphical models [24, 47, 141, 250]. Note that the self-explainability of these models is sometimes undermined by their complexity. For example, very complex tree or rule structures might sometimes be considered incomprehensible or unexplainable.

Some learning paradigms other than conventional models are also considered to be explainable, such as causal inference [197, 259] and the knowledge graph [345]. These methods are also expected to provide valuable inspiration to solve the problem of explainability for ML.

Beyond self-explainable ML models. Compared with black-box models, such as the DNN, conventional self-explainable models have poor performance on complex tasks, such as image classification and text comprehension. To achieve a compromise between explainability and performance, hybrid combinations of self-explainable models and black-box models have been proposed. A typical design involves embedding an explainable bottleneck model into a DNN. For example, previous research has embedded linear models and prototype selection to the DNN [16, 20, 73]. In the well-known class activation mapping [389], an average pooling layer at the end of a DNN can also be regarded as an explainable linear bottleneck. Attention mechanisms [30, 363] have also attracted recent interest and have been regarded as an explainable bottleneck in a DNN in some studies [79, 237]. However, this claim continues to be debated, because attention weights representing different explanations can produce similar final predictions [170, 355].

Post hoc model explanation. In addition to designing self-explainable models, understanding how a specific decision is made by black-box models is also an important problem. A major part of research on this problem has addressed the methodology of post hoc model explanation and proposed various approaches.

Explainer approximation aims to mimic the behavior of a given model with explainable models. This is also referred to as the global explanation of a model. Various approaches have

been proposed to approximate ML models, such as random forests [317, 392] and neural networks [28, 86, 393]. With the rise of deep learning in the past decade, explainer approximation on the DNN has advanced as the knowledge distillation problem on explainers such as trees [125, 384].

Feature importance has been a continually active area of research on explainability. A representative aspect uses local linear approximation to model the contribution of each feature to the prediction. **Local Interpretable Model-agnostic Explanation (LIME)** [279] and **SHapley Additive exPlanation (SHAP)** [225] are influential approaches that can be used for predictions on tabular data, computer vision, and NLP. Gradients can reflect how features contribute to the predicted outcome, and have drawn great interest in work on the explainability of the DNN [297, 305]. In NLP or **Computer Vision (CV)**, gradients or their variants are used to back-trace the decision of the model to the location of the most intimately related input, in the form of saliency maps and sentence highlights [250, 302, 314, 375].

Feature introspection aims to provide a semantic explanation of intermediate features. A representative aspect attaches an extra branch to a model to generate an explanatory outcome that is interpretable by a human. For example, in NLP-based reading comprehension, the generation of a supporting sentence serves as an explanatory task in addition to answer generation [332, 366]. In image recognition, part-template masks can be used to regularize feature maps to focus on local semantic parts [383]. Concept attribution [47] is another aspect that maps a given feature space to human-defined concepts. Similar ideas have been used in generative networks to gain control over attributes, such as the gender, age, and ethnicity, in a face generator [221].

Example-based explanation explains outcomes of the AI model by using the sample data. For example, an influential function was borrowed from robust statistics in Reference [193] to find the most influential data instance for a given outcome. Counterfactual explanation [13, 185, 341] works in a contrary way by finding the boundary case to flip the outcome. This helps users better understand the decision surface of the model.

3.2.4 Algorithmic Fairness. Methods to reduce bias in AI models during algorithm development can intervene before the data are fed into the model (pre-processing), when the model is being trained (in-processing), or into model predictions after it has been trained (post-processing).

Pre-processing methods. Aside from debiasing the data collection process, we can debias data before model training. Common approaches include the following:

Adjusting sample importance. This is helpful especially if debiasing the data collection is not sufficient or no longer possible. Common approaches include resampling [6], which involves selecting a subset of the data, reweighting [60], which involves assigning different importance values to data examples, and adversarial learning [229], which can be achieved through resampling or reweighting with the help of a trained model to find the offending cases.

Aside from helping to balance the classification accuracy, these approaches can be applied to balance the cost of classification errors to improve performance on certain groups [163] (e.g., for the screening of highly contagious and severe diseases, false negatives can be costlier than false positives; see *cost-sensitive learning* [321]).

Adjusting feature importance. Inadvertent correlation between features and sensitive variables can lead to unfairness. Common approaches of debiasing include representation transformation [61], which can help adjust the relative importance of features, and blinding [74], which omits features that are directly related to the sensitive variables.

Data augmentation. Besides making direct use of the existing data samples, it is possible to introduce additional samples that typically involves making changes to the available samples, including through perturbation and relabeling [60, 85].

In-processing methods. Pre-processing techniques are not guaranteed to have the desired effect during model training, because different models can leverage features and examples in different ways. This is where in-processing techniques can be helpful:

Adjusting sample importance. Similar to pre-processing methods, reweighting [195] and adversarial learning [68] can be used for in-processing, with the potential of either making use of the model parameters or predictions that are not yet fully optimized to more directly debias the model.

Optimization-related techniques. Alternatively, model fairness can be enforced more directly via optimization techniques. For instance, quantitative fairness metrics can be used as regularization [7] or constraints for the optimization of the model parameters [67].

Post-processing methods. Even if all precautions have been taken with regard to data curation and model training, the resulting models might still exhibit unforeseen biases. Post-processing techniques can be applied for debiasing, often with the help of auxiliary models or hyperparameters to adjust the model output. For instance, optimization techniques (e.g., constraint optimization) can be applied to train a smaller model to transform model outputs or calibrate model confidence [186]. Reweighting the predictions of multiple models can also help reduce bias [168].

3.2.5 Privacy Computing. Apart from privacy-preserving data processing methods, which were introduced in Section 3.1.2, another line of methods preserve data privacy during model learning. In this part, we briefly review the two popular categories of such algorithms: secure multi-party computing, and federated learning.

Secure Multi-party Computing (SMPC) deals with the task whereby multiple data owners compute a function, with the privacy of the data protected and no trusted third party serving as coordinator. A typical SMPC protocol satisfies properties of privacy, correctness, independence of inputs, guaranteed output delivery, and fairness [114, 387]. The garbled circuit is a representative paradigm for secure two-party computation [244, 367]. Oblivious transfer is among the key techniques. It guarantees that the sender does not know what information the receiver obtains from the transferred message. For the multi-party condition, secret sharing is one of the generic frameworks [181]. Each data instance is treated as a secret and split into several shares. These shares are then distributed to the multiple participating parties. The computation of the function value is decomposed into basic operations that are computed following the given protocol.

The use of SMPC in ML tasks has been studied in the context of both model-specific learning tasks, e.g., linear regression [128] and logistic regression [300], and generic model learning tasks [247]. Secure inference is the an emerging topic that tailors the SMPC for ML use. Its application to ML is as a service in which the server holds the model and clients hold the private data. To reduce the high costs of computation and communication of the SMPC, parameter quantization and function approximation were used together with cryptographic protocols in References [8, 32]. Several tools have been open sourced, such as MP2ML [48], CryptoSPN [330], CrypTFlow [200, 276], and CrypTen [188].

Federated learning (FL) was initially proposed as a secure scheme to collaboratively train an ML model on a data of user interactions with their devices [241]. It quickly gained extensive interest in academia and the industry as a solution to collaborative model training tasks by using data from multiple parties. It aims to address data privacy concerns that hinder ML algorithms from properly using multiple data sources. It has been applied to numerous domains, such as healthcare [282, 299] and finance [223].

Existing FL algorithms can be categorized into horizontal FL, vertical FL, and federated transfer learning algorithms [365]. Horizontal FL refers to the scenario in which each party has different samples but the samples share the same feature space. A training step is decomposed as to first compute optimization updates on each client and then aggregate them on a centralized server

without knowing the clients' private data [241]. Vertical FL refers to the setting in which all parties share the same sample ID space but have different features. Reference [148] used homomorphic encryption for vertical logistic regression-based model learning. In Reference [138], an efficient method of kernel learning was proposed. Federated transfer learning is applicable to the condition in which none of the parties overlaps in either the sample or the feature space [222]. The connection between FL and other research topics, such as multi-task learning, meta-learning, and fairness learning, has been discussed in Reference [180]. To expedite FL-related research and development, many open source libraries have been released, such as FATE, FedML [149], and Fedlearn-Algo [217].

3.3 Development

The manufacture of reliable products requires considerable effort in software engineering, and this is sometimes overlooked by AI developers. This lack of diligence, such as insufficient testing and monitoring, may incur long-term costs in the subsequent lifecycle of AI products (a.k.a. *technical debt* [296]). Software engineering in the stages of development and deployment has recently aroused wide concern as an essential condition for reliable AI systems [17, 203]. Moreover, various techniques researched for this stage can contribute to the trustworthiness of an AI system [17]. In this section, we survey the representative techniques.

3.3.1 Functional Testing. Inherited from the workflow of canonical software engineering, the testing methodology has drawn growing attention in the development of AI systems. In terms of AI trustworthiness, testing serves as an effective approach to certify that the system is fulfilling specific requirements. Recent research has explored to adapt functional testing to AI systems. This has been reviewed in the literature, such as References [164, 235, 381]. We describe two aspects of adaption from the literature that are useful to enhance the trustworthiness of an AI system.

Test criteria. Different from canonical software engineering where exact equity is tested between the actual and the expected outputs of a system, an AI system is usually tested by its predictive accuracy on a specific testing dataset. Beyond accuracy, various test criteria have been studied to further reflect and test more complex properties of an AI system. The concept of test coverage in software testing has been transplanted into DNN models [226, 260]. The name of a representative metric—neuron coverage [260]—figuratively illustrates that it measures the coverage of activated neurons in a DNN in analogy to code branches in canonical software testing. Such coverage criteria are effective for certifying the robustness of a DNN against adversarial attacks [226].

Test case generation. Human-annotated datasets are insufficient for thoroughly testing an AI system, and large-scale automatically generated test cases are widely used. Similar to canonical software testing, the problem to automatically generate the expected ground truth, known as the *oracle problem* [34], also occurs in the AI software testing scenario. The hand-crafted test case template is an intuitive but effective approach in applications of NLP [281]. Metamorphic testing is also a practical approach that converts the input/output pairs into new test cases. For example, [382] transfers images of road scenes taken in daylight to rainy images by using **Generative Adversarial Network (GAN)** as new test cases, and re-uses the original, invariant annotation to test an autonomous driving systems. These testing cases are useful for evaluating the generalization performance of an AI model. A similar methodology was adopted by adding adversarial patterns to normal images to test adversarial robustness [226]. Simulated environments are also widely used to test applications such as computer vision and reinforcement learning. We further review this topic in Section 3.3.3.

3.3.2 Performance Benchmarking. Unlike conventional software, the functionality of AI systems is often not easily captured in functional test alone. To ensure that systems are trustworthy

in terms of different aspects of interest, benchmarking (a.k.a. performance testing in software engineering) is often applied to ensure system performance and stability when these characteristics can be automatically measured.

Robustness is an important aspect of trustworthiness that is relatively amenable to automatic evaluation. References [88, 153] introduced a suite of black-box and white-box attacks to automatically evaluate the robustness of AI systems. It can potentially be performed as a sanity check before such systems are deployed to affect millions of users. Software fairness has also been a concern since conventional software testing [56, 127]. Criteria for AI systems have been studied to spot issues of unfairness by investigating the correlation among sensitive attributes, system outcomes, and the true label when applicable to well-designed diagnostic datasets [327]. Well-curated datasets and metrics have been proposed in the literature to evaluate performance on fairness metrics that are of interest for different tasks [40, 123, 307].

More recently, there has been growing interest in benchmarking explainability when models output explanations in NLP applications. For instance, Reference [238] asks crowd workers to annotate salient pieces of text that lead to their belief that the text is hateful or offensive, and examines how well model-predicted importance fits human annotations. Reference [93] instead introduces partial perturbations to the text for human annotators, and observes if the system's explanations match perturbations that change human decisions. In the meantime, Reference [267] reported that explainability benchmarking remains relatively difficult, because visual stimuli are higher dimensional and continuous.

3.3.3 Development by Simulation. While benchmarks serve to evaluate AI systems in terms of predictive behavior given static data, the behavior of many systems is deeply rooted in their interactions with the world. For example, benchmarking autonomous vehicle systems on static scenarios is insufficient to help us evaluate their performance on dynamic roadways. For these systems, simulation often plays an important role in ensuring their trustworthiness before deployment.

Robotics is a sub-fields of AI where simulations are most commonly used. Control systems for robots can be compared and benchmarked in simulated environments such as Gazebo [192], MuJoCo [324], and VerifAI [103]. Similarly, simulators for autonomous driving vehicles have been widely used, including CARLA [102], TORCS [359], CarSim [42], and PRESCAN [323]. These software platforms simulate the environment in which robots and vehicles operate as well as the actuation of controls on the simulated robots or cars. In NLP, especially conversational AI, simulators are widely used to simulate user behavior to test system ability and fulfill user needs by engaging in a dialog [205]. These simulators can help automatically ensure the performance of AI systems in an interactive environment and diagnose issues before their deployment.

Despite the efficiency, flexibility, and replicability afforded by software simulators, they often still fall short of perfectly simulating constraints faced by the AI system when deployed as well as environmental properties or variations in them. For AI systems that are deployed on embedded or otherwise boxed hardware, it is important to understand the system behavior when they are run on the hardware used in real-world scenarios. Hardware-in-the-loop simulations can help developers understand system performance when it is run on chips, sensors, and actuators in a simulated environment, and is particularly helpful for latency- and power-critical systems like autonomous driving systems [50, 54]. By taking real-world simulations a step further, one can also construct controlled real-world environments for fully integrated AI systems to roam around in (e.g., test tracks for self-driving cars with road signs and dummy obstacles). This provides more realistic measurements and assurances of performance before releasing such systems to users.

3.4 Deployment

After development, AI systems are deployed on realistic products, and interact with the environment and users. To guarantee that the systems are trustworthy, a number of approaches should be considered at the deployment stage, such as adding additional components to monitor anomalies, and developing specific human–AI interaction mechanisms for transparency and explainability.

3.4.1 Anomaly Monitoring. Anomaly monitoring has become a well-established methodology in software engineering. In terms of AI systems, the range of monitoring has been further extended to cover data outliers, data drifts, and model performance. As a keying safeguard for the successful operation of an AI system, monitoring provides the means to enhance the system's trustworthiness in multiple aspects. Some representative examples are discussed below.

Attack monitoring has been widely adopted in conventional SaaS, such as fraud detection [3] in e-commerce systems. In terms of the recent emerging adversarial attacks, detection and monitoring [243] of such attack inputs is also recognized as an important means to ensure system robustness. *Data drift monitoring* [268] provides important means to maintain the generalization of an AI system under concept change [394] caused by dynamic environment such as market change [289]. *Misuse monitoring* is recently also adopted in several cloud AI services [173] to avoid improper use such as unauthorized population surveillance or individual tracking by face recognition, which helps ensure the proper alignment of ethical values.

3.4.2 Human–AI Interaction. As an extension of **human–computer interaction (HCI)**, human–AI interaction has aroused wide attention in the AI industry [4, 18]. Effective human–AI interaction affects the trustworthiness of an AI system in multiple aspects. We briefly illustrate two topics.

User interface serves as the most intuitive factor affecting user experience. It is a major medium for an AI system to disclose its internal information and decision-making procedure to users, and thus has an important effect on the transparency and explainability of the system [301, 351]. Various approaches of interaction have been studied to enhance the explainability of AI, including the visualization of ML models [72] and interactive parameter-tuning [351]. In addition to transparency and explainability, the accessibility of the interface also significantly affects user experience of trustworthiness. AI-based techniques of interaction have enabled various new forms of human–machine interfaces, such as chatbots, audio speech recognition, and gesture recognition, and might result in accessibility problems for disabled people. Mitigating such unfairness has aroused concerns in recent research [179, 326].

Human intervention, such as by monitoring failure or participating in decisions [295], has been applied to various AI systems to compensate for limited performance. **Advanced Driving Assistance System (ADAS)** can be considered a typical example of systems involving human intervention, where the AI does the low-level driving work and the human makes the high-level decision. In addition to compensating for decision-making, human intervention provides informative supervision to train or fine-tune AI systems in many scenarios, such as the shadow mode [319] of autonomous driving vehicles. To minimize and make the best use of human effort in such interaction, efficient design of patterns of human–machine cooperation is an emerging topic in interdisciplinary work on HCI and AI and is referred to as *human-in-the-loop* or *interactive machine learning* [157] in the literature.

3.4.3 Fail-Safe Mechanisms. Considering the imperfection of current AI systems, it is important to avoid harm when the system fails in exceptional cases. By learning from conventional real-time automation systems, the AI community has realized that a fail-safe mechanism or fallback plan should be an essential part of the design of an AI system if its failure can cause harm or loss.

This mechanism is also emerging as an important requirement in recent AI guidelines, such as Reference [9]. The fail-safe design has been observed in multiple areas of robotics in the past few years. In the area of **Unmanned Aerial Vehicle (UAV)**, the fail-safe algorithm has been studied for a long time to avoid frequent collision of quadrocopters [126] and to ensure safe landing upon system failure [252]. In autonomous driving where safety is critical, a fail-safe mechanism like standing still has become an indispensable component in **Advanced Driver-Assistant System (ADAS)** products [160], and is being researched at a higher level of automation [230].

3.4.4 Hardware Security. AI systems are widely deployed on various hardware platforms to cope with the diverse scenarios, ranging from servers in computing centers to cellphones and embedded systems. Attacks on OS and hardware lead to new risks, such as data tampering or stealing, and threaten the robustness, security, and privacy of AI systems. Various approaches have been studied to address this new threat [364]. From the perspective of hardware security, the concept of a **trusted execution environment (TEE)** is a recent representative technique that has been adopted by many hardware manufacturers [287]. The general mechanism of the TEE is to provide a secure area for the data and the code. This area is not interfered with by the standard OS such that the protected program cannot be attacked. ARM processors support TEE implementation using the TrustZone design [264]. They simultaneously run a secure OS and a normal OS on a single core. The secure part provides a safe environment for sensitive information. The Intel Software Guard Extensions implement the TEE by hardware-based memory encryption [240]. Its enclave mechanism allows for the allocation of protected memory to hold private information. Such security mechanisms have been used to protect sensitive information like biometric ID and financial account passwords and are applicable to other AI use cases.

3.5 Management

AI practitioners such as researchers and developers have studied various techniques to improve AI trustworthiness in the aforementioned stages of the data, algorithm, development, and deployment stages. Beyond these concrete approaches, appropriate management and governance provide a holistic guarantee that trustworthiness is consistently aligned throughout the lifecycle of a AI system. In this section, we introduce several executable approaches that facilitate the AI community to improve management and governance on AI trustworthiness.

3.5.1 Documentation. Conventional software engineering has accumulated a wealth of experience in leveraging documentation to assist development. Representative documentation types include requirement documents, product design documents, architecture documents, code documents, and test documents [11]. Beyond conventional software engineering, multiple new types of documents have been proposed to adapt to the ML training and testing mechanisms. Their scope may include the purposes and characteristics of the model [246], datasets [41, 129, 156], and services [22]. As mentioned in Sections 2.3.2 and 2.7, documentation is an effective and important approach to enhance the system's transparency and accountability by tracking, guiding, and auditing its entire lifecycle [272] and serves as a cornerstone of building a trustworthy AI system.

3.5.2 Auditing. With lessons learned from safety-critical industries, e.g., finance and aerospace, auditing has been recently recognized as an effective mechanism to examine whether an AI system complies with specific principles [58, 356]. In terms of the position of auditors, the auditing process can be categorized as internal or external. Internal auditing enables self-assessment and iterative improvement for manufacturers to follow the principles of trustworthiness. It can cover the lifecycle of the system without the leakage of trade secrets [272]. However, external auditing by independent parties is more effective in gaining public trust [58].

Auditing might involve the entire or selective parts of the lifecycle of an AI system. A comprehensive framework of internal auditing can be found in Reference [272]. The means of auditing might include interviews, documented artifacts, checklists, code review, testing, and impact assessment. For example, documentation like product requirement documents, model cards [246], and datasheets [129] serve as important references to understand the principle alignment during development. Checklists are widely used as a straightforward qualitative approach to evaluate fairness [228], transparency [292], and reproducibility [263]. Quantitative testing also serves as a powerful approach and has been successfully executed to audit fairness in, for example, the Gender Shade study [58]. Inspired by the EU's **European Union's Data Protection Impact Assessment (DPIA)**, the concept of Algorithmic Impact Assessment has been proposed to evaluate the claims of trustworthiness and discover negative effects [277]. Besides the above representatives, designs of approaches to algorithmic auditing can be found in References [290, 356].

3.5.3 Cooperation and Information Sharing. As shown in Figure 2, the establishment of trustworthy AI requires cooperation between stakeholders. From the perspective of industry, cooperation with academia enables the fast application of new technology to enhance the performance of the product and reduce the risk posed by it. Cooperation with regulators certifies the products as appropriately following the principles of trustworthiness. Moreover, cooperation between industrial enterprises helps address consensus-based problems, such as data exchange, standardization, and ecosystem building [27]. Recent practices of AI stakeholders have shown the efficacy of cooperation in various dimensions. We summarize these practices in the following aspects below.

Collaborative research and development. Collaboration has been a successful driving force in the development of AI technology. To promote research on AI trustworthiness, stakeholders are setting-up various forms of collaboration, such as research workshops on trustworthy AI and cooperative projects like DARPA XAI [144].

Trustworthy data exchange. The increasing business value of data raises the demand to exchange them across companies in various scenarios (e.g., the medical AI system in Section 2.6). Beyond privacy-based computing techniques, the cooperation between data owners, technology providers, and regulators is making progress in establishing an ecosystem of data exchange, and solving problems such as data pricing and data authorization.

Cooperative development of regulation. Active participation in the development of standards and regulations serves as an important means for academia, the industry, and regulators to align their requirements and situations.

Incident sharing. The AI community has recently recognized incident sharing as an effective approach to highlight and prevent potential risks to AI systems [57]. The AI Incident Database [91] provides an inspiring example for stakeholders to share negative AI incidents so that the industry can avoid similar problems.

3.6 TrustAIOps: A Continuous Workflow toward Trustworthiness

The problem of trustworthy AI arises from the fast development of AI technology and its emerging applications. AI trustworthiness is not a well-studied static bar to reach by some specific solutions. The establishment of trustworthiness is a dynamic procedure. We have witnessed the evolution of different dimensions of trustworthiness over the past decade [178]. For example, research on adversarial attacks has increased concerns regarding adversarial robustness. The applications of safety-critical scenarios have rendered more stringent the requirements of accountability of an AI system. The development of AI research, the evolution of the forms of AI products, and the changing perspectives of society imply the continual reformulation of the requirements of and solutions to trustworthiness. Therefore, we argue that beyond the requirements of an AI product, the AI

industry should consider trustworthiness as an ethos of its operational routine and be prepared to continually enhance the trustworthiness of its products.

The constant enhancement of AI trustworthiness positions requirements on a new workflow for the AI industry. Recent studies on industrial AI workflow extend the mechanism of DevOps [36] to MLOps [233], to enable improvements in ML products. The concept of DevOps has been adopted in modern software development to continually deploy software features and improve their quality. MLOps [233] and its variants, such as ModelOps [165] and SafetyOps [303], extend DevOps to cover the ML lifecycle of data preparation, training, validation, and deployment, in its workflow. The workflow of MLOps provides a start point to build the workflow for trustworthy AI. By integrating the ML lifecycle, MLOps connects research, experimentation, and product development to enable the rapid leveraging of the theoretical development of trustworthy AI. A wealth of toolchains of MLOps have been released recently to track AI artifacts such as data, model, and meta-data to increase the accountability and reproducibility of products [165]. Recent research has sought to extend MLOps to further integrate trustworthiness into the AI workflow. For example, [303] extended MLOps with safety engineering as SafetyOps for autonomous driving.

As we have illustrated in this section, building trustworthiness requires the continual and systematic upgrade of the AI lifecycle. By extending MLOps, we summarize this upgrade of practices as a new workflow, *TrustAIOps*, which focuses on imposing the requirements of trustworthiness over the entire AI lifecycle. This new workflow contains the following properties:

- **Close collaboration between inter-disciplinary roles.** Building trustworthy AI requires organizing different roles, such as ML researchers, software engineers, safety engineers, and legal experts. Close collaboration mitigates the gap in knowledge between forms of expertise (e.g., Reference [208], cf., Sections 3.5.3 and A.2).
- **Aligned principles of trustworthiness.** The risk of untrustworthiness exists in every stage in the lifecycle of an AI system. Mitigating such risks requires that all stakeholders in the AI industry be aware of and aligned with unified trustworthy principles (e.g., Reference [301], cf., Section A.2).
- **Extensive management of artifacts.** An industrial AI system is built upon various artifacts such as data, code, models, configuration, product design, and operation manuals. The elaborate management of these artifacts helps assess risk and increases reproducibility and auditability (cf., Section 3.5.1).
- **Continuous feedback loops.** Classical continuous integration and continuous development (CI/CD) workflows provide effective mechanisms to improve the software through feedback loops. In a trustworthy AI system, these feedback loops should connect and iteratively improve the five stages of its lifecycle, i.e., data, algorithm, development, deployment, and management (e.g., References [272, 310]).

The evolution of the industrial workflow of AI is a natural reflection of the dynamic procedure to establish its trustworthiness. By systematically organizing stages of the AI lifecycle and inter-disciplinary practitioners, the AI industry is able to understand the requirements of trustworthiness from various perspectives, including technology, law, and society, and deliver continual improvements.

4 CONCLUSION, CHALLENGES, AND OPPORTUNITIES

In this survey, we outlined the key aspects of trustworthiness that we think are essential to AI systems. We introduced how AI systems can be evaluated and assessed on each of these aspects, and reviewed current efforts in this direction in the industry. We further proposed a systematic approach to consider these aspects of trustworthiness in the entire lifecycle of real-world AI

systems, which offers recommendations for every step of the development and use of these systems. We recognize that fully adopting this systematic approach to build trustworthy AI systems requires that practitioners embrace the concepts underlying the key aspects that we have identified. More importantly, it requires a shift of focus from *performance-driven* AI to *trust-driven* AI. In the short run, this shift will inevitably involve side-effects, such as longer learning time, slowed development, and/or increased cost to build AI systems. However, we encourage practitioners to focus on the long-term benefits of gaining the trust of all stakeholders for the sustained use and development of these systems. In this section, we conclude by discussing some of the open challenges and potential opportunities in the future development of trustworthy AI.

4.1 AI Trustworthiness as Long-term Research

Our understanding of AI trustworthiness is far from complete or universal and will inevitably evolve as we develop new AI technologies and understand their societal impact more clearly. This procedure requires long-term research in multiple key areas of AI. In this section, we discuss several open questions that we think are crucial to address for the future development of AI trustworthiness.

4.1.1 Immaturity of Approaches to Trustworthiness. As mentioned in Section 2, several aspects of AI trustworthiness, such as explainability and robustness, address the limitation of current AI technologies. Despite wide interest in AI research, satisfactory solutions are still far from reach.

Consider explainability as an example. Despite being an active field of AI research, it remains poorly understood. Both current explanatory models and post hoc model explanation techniques share a few common issues, e.g., (1) the explanation is fragile to perturbations [130], (2) the explanation is not always consistent with human interpretation [47], and (3) it is difficult to judge if the explanation is correct or faithful [250]. These problems pose important questions in the study of explainability and provide valuable directions of research in theoretical research on AI.

Another example is robustness. The arms race between adversarial attack and defense reflects the immaturity of our understanding of the robustness of AI. As in other areas of security, attacks evolve along with the development of defenses. Conventional adversarial training [134] has been shown to be easily fooled by subsequently developed attacks [328]. The corresponding defense [328] is later shown to be vulnerable against new attacks [99]. This not only requires that practitioners be agile in adopting defensive techniques to mitigate the risk of new attacks in a process of long-term and continual development but also poses long-term challenges to theoretical research [270].

4.1.2 Frictional Impact of Trustworthy Aspects. As we have shown in Section 2, there are a wealth of connections and support between different aspects of trustworthiness. However, research has shown that there are frictions or tradeoffs between these aspects in some cases, which we review here.

Increased transparency improves trust in AI systems through information disclosure. However, disclosing inappropriate information might increase potential risks. For example, excessive transparency on datasets and algorithms might leak private data and commercial intellectual property. Disclosure of detailed algorithmic mechanisms can also lead to the risk of targeted hacking [12]. However, an inappropriate explanation might also cause users to overly rely on the system and follow wrong decisions of AI [311]. Therefore, the extent of transparency of an AI system should be specified carefully and differently for the roles of public users, operators, and auditors.

From an algorithmic perspective, the effects of different objectives of trustworthiness on model performance remain insufficiently understood. Adversarial robustness increases the model's generalizability and reduces overfitting, but tends to negatively impact its overall accuracy [331, 380].

A similar loss of accuracy occurs in explainable models [47]. Besides this trust–accuracy tradeoff, algorithmic friction exists between the dimensions of trustworthiness. For example, adversarial robustness and fairness can negatively affect each other during training [284, 361]. Furthermore, studies on fairness and explainability have shown several approaches to explanation to be unfair [98].

These frictional effects suggest that AI trustworthiness cannot be achieved through hillclimbing on a set of disjoint criteria. Compatibility should be carefully considered when integrating multiple requirements into a single system. Recent studies [361, 380] provide a good reference to start with.

4.1.3 Limitations in Current Evaluations of Trustworthiness. Repeatable and quantitative measurements are the cornerstone of scientific and engineering progress. However, despite increasing research interest and efforts, the quantification of many aspects of AI trustworthiness remains elusive. Of the various aspects that we have discussed in this article, the explainability, transparency, and accountability of AI systems are still seldom evaluated quantitatively, which makes it difficult to accurately compare systems. Developing good methods of quantitative evaluation for these desiderata, we believe, will be an important first step in research on these aspects of AI trustworthiness as a scientific endeavor, rather than a purely philosophical one.

4.1.4 Challenges and Opportunities in the Era of Large-scale Pre-trained Models. Large scale pre-trained models have brought dramatic breakthroughs for AI. They not only show the potential to a more general form of AI [234] but also bring new challenges and opportunities to the establishment of trustworthy AI. One of the most important properties of a large scale pre-trained model is the ability to transfer its learned knowledge to new tasks in a manner of few-shot or zero-shot learning [55]. This largely fulfills people's requirement on the generalization of AI, and is recognized to hold great value in commercial applications. However, the risks of the large scale pre-trained models to be untrustworthy have been revealed by recent studies. For example, the training procedure is known to be costly and difficult to reproduce for third-parties, as mentioned in Section 2.2. Most downstream tasks have to directly adapt the pre-trained model without auditing its entire lifecycle. This business model poses a risk that downstream users might be affected by any biases presented in these models [234]. Privacy leakage is another issue that is recently revealed. Some pre-trained models are reported to output training text data containing private user information such as addresses [63].

The development and application of the large scale pre-trained models is accelerating. To guarantee that this advancement will benefit the society without causing new risks, it is worthwhile for both academia and the industry to carefully study its potential impact in the perspective of AI trustworthiness.

4.2 End-User Awareness of the Importance of AI Trustworthiness

Other than developers and providers of AI systems, end-users are an important yet opt-ignored group of stakeholders.

Besides educating the general public about the basic concepts of AI trustworthiness, developers should consider how it can be presented to users to deliver hands-on experiences of trustworthy AI systems. One positive step in this direction involves demonstrating system limitations (e.g., Google Translate displays translations in multiple gendered pronouns when the input text is ungendered) or explainable factors used for system prediction (e.g., recommendation systems can share user traits that are used in curating ads, like "female aged 20–29" and "interested in technology"). Taking this a step further, we believe that it would enable the user to directly control these factors (e.g., user traits) counterfactually, and judge for themselves whether the system is fair, robust, and trustworthy.

Finally, we recognize that not all aspects of trustworthiness can be equally easily conveyed to end-users. For instance, the impact of privacy preservation or transparency cannot be easily demonstrated to the end-user when AI systems are deployed. We believe that media coverage and government regulations regarding these aspects can be very helpful in raising public awareness.

4.3 Inter-disciplinary and International Cooperation

An in-depth understanding of AI trustworthiness involves not only the development of better and newer AI technologies, but also requires us to better understand the interactions between AI and human society. We believe this calls for collaboration across various disciplines that reach far beyond computer science. First, AI practitioners should work closely with domain experts whenever AI technologies are deployed to the real world and has impacts on people, e.g., in medicine, finance, transportation, and agriculture. Second, AI practitioners should seek advice from social scientists to better understand the (often unintended) societal impacts of AI and work together to remedy them, e.g., the impact of AI-automated decisions, job displacement in AI-impacted sectors, and the effect of the use of AI systems in social networks. Third, AI practitioners should carefully consider how the technology is presented to the public as well as inter-disciplinary collaborators, and make sure to communicate the known limitations of AI systems honestly and clearly.

In the meantime, the development of trustworthy AI is by no means a unique problem of any single country, nor does the potential positive or negative effect of AI systems respect geopolitical borders. Despite the common portrayal of AI as a race between countries (see Section 3.2 of Reference [145]), technological advances in the climate of increased international collaboration are far from a zero-sum game. It not only allows us to build better technological solutions by combining diverse ideas from different backgrounds but also helps us better serve the world's population by recognizing our shared humanity as well as our unique differences. We believe that tight-knit inter-disciplinary and international cooperation will serve as the bedrock for rapid and steady developments in trustworthy AI technology, which will in turn benefit humanity at large.

ACKNOWLEDGMENTS

The authors would like to thank Yanqing Chen, Jing Huang, Shuguang Zhang, and Liping Zhang for their valuable suggestions. We also thank Yu He, Wenhan Xu, Xinyuan Shan, Chenliang Wang, Peng Liu, Jingling Fu, Baicun Zhou, Hongbao Tian, Qili Wang and Ermo Hua for their assistance.

REFERENCES

[1] David Krueger, Nicolas Ballas, Stanislaw Jastrzebski, Devansh Arpit, Maxinder S. Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, and Aaron C. Courcille. Deep nets don't learn via memorization. In *ICLR (Workshop)*.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 308–318.

[3] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. 2016. Fraud detection system: A survey. *J. Netw. Comput. Appl.* 68 (2016), 90–113.

[4] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[5] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[6] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.* 54, 1 (2018), 95–122.

[7] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1418–1426.

[8] Nitin Agrawal, Ali Shahin Shamsabadi, Matt J. Kusner, and Adrià Gascón. 2019. QUOTIENT: Two-party secure neural network training and prediction. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1231–1247.
[9] AI HLEG, European Commission. 2021. Ethics guidelines for trustworthy AI.
[10] AI HLEG, European Commission. 2020. Assessment list for trustworthy artificial intelligence (altai) for self-assessment.
[11] Alberto Aimar. 1998. Introduction to software documentation.
[12] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access* 6 (2018), 14410–14430.
[13] Arjun R. Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. 2022. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience* 25, 1 (2022), 103581.
[14] Dalal Al-Azizy, David Millard, Iraklis Symeonidis, Kieron O'Hara, and Nigel Shadbolt. 2015. A literature survey and classifications on data deanonymisation. In *Proceedings of the International Conference on Risks and Security of Internet and Systems*. Springer, 36–51.
[15] Mohammed AlShamsi, Said A. Salloum, Muhammad Alshurideh, and Sherief Abdallah. 2021. Artificial intelligence and blockchain for transparency in governance. In *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications*. Springer, 219–230.
[16] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems* 31 (2018).
[17] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP'19)*. IEEE, 291–300.
[18] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the Chi Conference on Human Factors in Computing Systems*. 1–13.
[19] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. arXiv:1606.06565. Retrieved from https://arxiv.org/abs/1606.06565.
[20] Plamen Angelov and Eduardo Soares. 2020. Towards explainable deep neural networks (xDNN). *Neural Networks* 130 (2020), 185–194.
[21] Apple. 2017. Differential Privacy. Retrieved February 20, 2021 from https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.
[22] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K. Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.* 63, 4/5 (2019), 6–1.
[23] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 233–242.
[24] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fus.* 58 (2020), 82–115.
[25] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv:1909.03012. Retrieved from https://arxiv.org/abs/1909.03012.
[26] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput. Surv.* 54, 5 (2021), 1–39.
[27] Amanda Askell, Miles Brundage, and Gillian Hadfield. 2019. The role of cooperation in responsible AI development. arXiv:1907.04534. Retrieved from https://arxiv.org/abs/1907.04534.
[28] M. Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. 2012. Reverse engineering the neural networks for rule extraction in classification problems. *Neural Process. Lett.* 35, 2 (2012), 131–150.
[29] Serkan Ayvaz and Salih Cemil Cetin. 2019. Witness of things: Blockchain-based distributed decision record-keeping system for autonomous vehicles. *Int. J. Intell. Unman. Syst.* (2019).
[30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR'15)*.
[31] Josef Baker-Brunnbauer. 2021. Taii framework for trustworthy ai systems. *ROBONOMICS: The Journal of the Automated Economy*, 2 (2021), 17–17.

[32] Marshall Ball, Brent Carmer, Tal Malkin, Mike Rosulek, and Nichole Schimanski. 2019. Garbled neural networks are practical. *Cryptology ePrint Archive*.
[33] Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, and Jaehoon Amir Safavi. 2017. Mitigating poisoning attacks on machine learning models: A data provenance based approach. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 103–110.
[34] Earl T. Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo. 2014. The oracle problem in software testing: A survey. *IEEE Trans. Softw. Eng.* 41, 5 (2014), 507–525.
[35] Peter L. Bartlett and Shahar Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* 3 (November 2002), 463–482.
[36] Len Bass, Ingo Weber, and Liming Zhu. 2015. *DevOps: A Software Architect's Perspective*. Addison-Wesley Professional.
[37] Ghazaleh Beigi and Huan Liu. 2020. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Trans. Data Sci.* 1, 1 (2020), 1–38.
[38] Wei Zhang, Yang Yu, and Bowen Zhou. 2015. Structured Memory for Neural Turing Machines. arXiv: 1510.03931. Retrieved from https://arxiv.org/abs/1510.03931.
[39] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* 116, 32 (2019), 15849–15854.
[40] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, 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), 4–1.
[41] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguist.* 6 (2018), 587–604.
[42] Rahim F. Benekohal and Joseph Treiterer. 1988. CARSIM: Car-following model for simulation of traffic in normal and stop-and-go conditions. *Transp. Res. Rec.* 1194 (1988), 99–111.
[43] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50, 1 (2021), 3–44.
[44] Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* 94, 4 (2004), 991–1013.
[45] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. 2019. A survey of black-box adversarial attacks on computer vision models. arXiv:1912.01667. Retrieved from https://arxiv.org/abs/1912.01667.
[46] Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martinez, and David Sanchez. 2020. Machine learning explainability via microaggregation and shallow decision trees. *Knowl.-Bas. Syst.* 194 (2020), 105532.
[47] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and survey of explanation methods for black box models. arXiv:2102.13076. Retrieved from https://arxiv.org/abs/2102.13076.
[48] Fabian Boemer, Rosario Cammarota, Daniel Demmler, Thomas Schneider, and Hossein Yalame. 2020. MP2ML: A mixed-protocol machine learning framework for private inference. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. 1–10.
[49] Miranda Bogen and Aaron Rieke. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.
[50] Thomas Bokc, Markus Maurer, and Georg Farber. 2007. Validation of the vehicle in the loop (vil); a milestone for the simulation of driver assistance systems. In *Proceedings of the IEEE Intelligent Vehicles Symposium*. IEEE, 612–617.
[51] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. 2019. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3240–3247.
[52] Neal E. Boudette. "It Happened So Fast": Inside a Fatal Tesla Autopilot Accident. Retrieved from https://www.nytimes.com/2021/08/17/business/tesla-autopilot-accident.html.
[53] Timo Breuer, Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, Philipp Schaer, and Ian Soboroff. 2020. How to measure the reproducibility of system-oriented IR experiments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 349–358.
[54] Craig Brogle, Chao Zhang, Kai Li Lim, and Thomas Bräunl. 2019. Hardware-in-the-loop autonomous driving simulation without real-time constraints. *IEEE Trans. Intell. Vehic.* 4, 3 (2019), 375–384.
[55] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33 (2020), 1877–1901.
[56] Yuriy Brun and Alexandra Meliou. 2018. Software fairness. In *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 754–759.
[57] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:2004.07213. Retrieved from https://arxiv.org/abs/2004.07213.

[58] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.
[59] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *Proceedings of the IEEE 13th International Conference on Data Mining*. IEEE, 71–80.
[60] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Minl. Knowl. Discov.* 21, 2 (2010), 277–292.
[61] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3995–4004.
[62] Rosario Cammarota, Matthias Schunter, Anand Rajan, Fabian Boemer, Ágnes Kiss, Amos Treiber, Christian Weinert, Thomas Schneider, Emmanuel Stapf, Ahmad-Reza Sadeghi, et al. 2020. Trustworthy AI inference systems: An industry research view. arXiv:2008.04449. Retrieved from https://arxiv.org/abs/2008.04449.
[63] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security'21)*. 2633–2650.
[64] Nicholas Carlini and David Wagner. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 3–14.
[65] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 39–57.
[66] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. arXiv:2010.04053. Retrieved from https://arxiv.org/abs/2010.04053.
[67] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.
[68] L. Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. arXiv:1901.10443. Retrieved from https://arxiv.org/abs/1901.10443.
[69] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. arXiv:1810.00069. Retrieved from https://arxiv.org/abs/1810.00069.
[70] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep learning for anomaly detection: A survey. arXiv:1901.03407. Retrieved from https://arxiv.org/abs/1901.03407.
[71] Patrick P. K. Chan, Zhi-Min He, Hongjiang Li, and Chien-Chang Hsu. 2018. Data sanitization against adversarial label contamination based on data complexity. *Int. J. Mach. Learn. Cybernet.* 9, 6 (2018), 1039–1052.
[72] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 713–756.
[73] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. 2019. This looks like that: Deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[74] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Adv. Neural Inf. Process. Syst.* 31 (2018).
[75] Rui Chen, Benjamin Fung, Philip S. Yu, and Bipin C. Desai. 2014. Correlated network data publication via differential privacy. *VLDB J.* 23, 4 (2014), 653–676.
[76] Rui Chen, Noman Mohammed, Benjamin C. M. Fung, Bipin C. Desai, and Li Xiong. 2011. Publishing set-valued data via differential privacy. *Proc. VLDB Endow.* 4, 11 (2011), 1087–1098.
[77] Yizheng Chen, Shiqi Wang, Yue Qin, Xiaojing Liao, Suman Jana, and David Wagner. 2021. Learning security classifiers with verified global robustness properties. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 477–494.
[78] Yu Cheng, Mo Yu, Xiaoxiao Guo, and Bowen Zhou. 2019. Few-shot learning with meta metric learners. In *NIPS 2017 Workshop on Meta-Learning*.
[79] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv. Neural Inf. Process. Syst.* 29 (2016).
[80] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
[81] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the International Conference on Management of Data*. 2201–2206.
[82] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. 2017. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*. PMLR, 854–863.

[83] Edmund M. Clarke and Jeannette M. Wing. 1996. Formal methods: State of the art and future directions. *ACM Comput. Surv.* 28, 4 (1996), 626–643.
[84] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023. Retrieved from https://arxiv.org/abs/1808.00023.
[85] Bo Cowgill and Catherine Tucker. 2017. Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms.*
[86] Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Adv. Neural Inf. Process. Syst.* 8 (1995).
[87] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. 2008. Casting out demons: Sanitizing training data for anomaly sensors. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'08)*. IEEE, 81–95.
[88] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: A standardized adversarial robustness benchmark. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
[89] Luiz Marcio Cysneiros and Vera Maria Benjamim Werneck. 2009. An initial analysis on how software transparency and trust influence each other. In *Proceedings of the Workshop on Requirements Engineering (WER'09)*. Citeseer.
[90] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. arXiv:2012.08630. Retrieved from https://arxiv.org/abs/2012.08630.
[91] David Dao. 2020. Awful AI. Retrieved from https://github.com/daviddao/awful-ai.
[92] Jacob Devlin, Ming-Wei Chang Chang, Lee Kenton, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
[93] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. *Trans. Assoc. Comput. Linguist.* (2020).
[94] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.
[95] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. Retrieved from https://www.documentcloud.org/documents/2998391-ProPublica-CommentaryFinal-070616.html.
[96] Donna N. Dillenberger, Petr Novotny, Qi Zhang, Praveen Jayachandran, Himanshu Gupta, Sandeep Hans, Dinesh Verma, Shreya Chakraborty, J. J. Thomas, M. M. Walli, et al. 2019. Blockchain analytics and artificial intelligence. *IBM J. Res. Dev.* 63, 2/3 (2019), 5–1.
[97] Ahmet Emir Dirik, Hüsrev Taha Sencar, and Nasir Memon. 2014. Analysis of seam-carving-based anonymization of images against PRNU noise pattern-based source attribution. *IEEE Trans. Inf. Forens. Sec.* 9, 12 (2014), 2277–2290.
[98] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.
[99] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4312–4321.
[100] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Shieber, James Waldo, David Weinberger, et al. 2018. Accountability of AI under the law: The role of explanation. In *Privacy Law Scholars Conference*.
[101] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'18)*. IEEE, 0210–0215.
[102] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on Robot Learning*. PMLR, 1–16.
[103] Tommaso Dreossi, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A. Seshia. 2019. Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *Proceedings of the International Conference on Computer Aided Verification*. Springer, 432–442.
[104] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4, 1 (2018), eaao5580.
[105] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Trans. Technol. Soc.* 1, 2 (2020), 89–103.
[106] Chris Drummond. 2009. Replicability is not reproducibility: Nor is it good science.

[107] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.
[108] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.
[109] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2016. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confid.* 7, 3 (2016), 17–51.
[110] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.
[111] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS One* 6, 12 (2011), e28071.
[112] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning?. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 201–208.
[113] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1054–1067.
[114] David Evans, Vladimir Kolesnikov, Mike Rosulek, et al. 2018. A pragmatic introduction to secure multi-party computation. *Found. Trends Priv. Secur.* 2, 2-3 (2018), 70–246.
[115] Exforsys. 2011. What Is Monkey Testing. Retrieved July 9, 2021 from http://www.exforsys.com/tutorials/testing-types/monkey-testing.html.
[116] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. 2018. Adversarial vulnerability for any classifier. *Adv. Neural Inf. Process. Syst.* 31 (2018).
[117] 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.
[118] Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2018. Overview of CENTRE@ CLEF 2018: A first tale in the systematic reproducibility realm. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 239–246.
[119] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.
[120] Sam Fletcher and Md Zahidul Islam. 2019. Decision tree classification with differential privacy: A survey. *ACM Comput. Surv.* 52, 4 (2019), 1–33.
[121] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nat. Mach. Intell.* 1, 6 (2019), 261–262.
[122] Luciano Floridi and Josh Cowls. 2019. A unified framework of five principles for AI in society. *Harv. Data Sci. Review* 1, 1 (2019).
[123] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.
[124] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. 2001. *The Elements of Statistical Learning*, Vol. 1. Springer. New York.
[125] Nicholas Frosst and Geoffrey Hinton. 2017. Distilling a neural network into a soft decision tree. In *CEX Workshop at AI*IA 2017 Conference*.
[126] Changhong Fu, Miguel A. Olivares-Mendez, Ramon Suarez-Fernandez, and Pascual Campoy. 2014. Monocular visual-inertial SLAM-based collision avoidance strategy for fail-safe UAV using fuzzy logic controllers. *J. Intell. Robot. Syst.* 73, 1 (2014), 513–533.
[127] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*. 498–510.
[128] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. 2017. Privacy-preserving distributed linear regression on high-dimensional data. *Proc. Priv. Enhanc. Technol.* 2017, 4 (2017), 345–364.
[129] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[130] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.
[131] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 580–587.

[132] John R. Goodall, Eric D. Ragan, Chad A. Steed, Joel W. Reed, G. David Richardson, Kelly M. T. Huffer, Robert A. Bridges, and Jason A. Laska. 2018. Situ: Identifying and explaining suspicious behavior in networks. *IEEE Trans. Vis. Comput. Graph.* 25, 1 (2018), 204–214.
[133] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Machine learning basics. *Deep Learn.* 1 (2016), 98–164.
[134] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.
[135] Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38, 3 (2017), 50–57.
[136] Google. 2020. Responsible AI with TensorFlow. Retrieved February 20, 2021 https://blog.tensorflow.org/2020/06/responsible-ai-with-tensorflow.html.
[137] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[138] Bin Gu, Zhiyuan Dang, Xiang Li, and Heng Huang. 2020. Federated doubly stochastic kernel learning for vertically partitioned data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2483–2493.
[139] Shixiang Gu and Luca Rigazio. 2015. Towards deep neural network architectures robust to adversarial examples. In *3rd International Conference on Learning Representations (ICLR'15), San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.
[140] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. arXiv:1805.10820. Retrieved from https://arxiv.org/abs/1805.10820.
[141] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5 (2018), 1–42.
[142] Odd Erik Gundersen, Yolanda Gil, and David W. Aha. 2018. On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Mag.* 39, 3 (2018), 56–68.
[143] Odd Erik Gundersen and Sigbjorn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[144] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* 40, 2 (2019), 44–58.
[145] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* 30, 1 (2020), 99–120.
[146] Karen Hao. 2019. AI Is Sending People to Jail—And Getting It Wrong. Retrieved February 20, 2021 from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/. Accessed: 2021-02-20.
[147] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Adv. Neur. Inf. Process. Syst.* 29 (2016).
[148] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. 2017. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv:1711.10677. Retrieved from https://arxiv.org/abs/1711.10677.
[149] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. 2020. Fedml: A research library and benchmark for federated machine learning. arXiv:2007.13518. Retrieved from https://arxiv.org/abs/2007.13518.
[150] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.
[151] Yeye He and Jeffrey F. Naughton. 2009. Anonymization of set-valued data via top-down, local generalization. *Proc. VLDB Endow.* 2, 1 (2009), 934–945.
[152] Matthias Hein and Maksym Andriushchenko. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Adv. Neur. Inf. Process. Syst.* 30 (2017).
[153] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.
[154] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A survey on provenance: What for? What form? What from? *VLDB J.* 26, 6 (2017), 881–906.
[155] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608. Retrieved from https://arxiv.org/abs/1812.04608.
[156] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. In *Data Protection and Privacy, Volume 12: Data Protection and Democracy*, Vol. 3. 1–1.

[157] Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inf.* 3, 2 (2016), 119–131.
[158] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. 2021. Federated robustness propagation: Sharing adversarial robustness in federated learning. arXiv:2106.10196. Retrieved from https://arxiv.org/abs/2106.10196.
[159] Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. 2014. Collaborative search log sanitization: Toward differential privacy and boosted utility. *IEEE Trans. Depend. Sec. Comput.* 12, 5 (2014), 504–518.
[160] Markus Hörwick and Karl-Heinz Siedersberger. 2010. Strategy and architecture of a safety concept for fully automatic and autonomous driving assistance systems. In *Proceedings of the IEEE Intelligent Vehicles Symposium*. IEEE, 955–960.
[161] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Sci. Eng. Ethics* 24, 5 (2018), 1521–1536.
[162] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. 2021. Model complexity of deep learning: A survey. *Knowl. Inf. Syst.* 63, 10 (2021), 2585–2619.
[163] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5375–5384.
[164] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.* 37 (2020), 100270.
[165] Waldemar Hummer, Vinod Muthusamy, Thomas Rausch, Parijat Dube, Kaoutar El Maghraoui, Anupama Murthi, and Punleuk Oum. 2019. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *Proceedings of the IEEE International Conference on Cloud Engineering (IC2E'19)*. IEEE, 113–120.
[166] IBM. Trusting AI. Retrieved February 20, 2021 from https://www.research.ibm.com/artificial-intelligence/trusted-ai/.
[167] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. PMLR, 448–456.
[168] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2019. Fae: A fairness-aware ensemble framework. In *Proceedings of the IEEE International Conference on Big Data (Big Data'19)*. IEEE, 1375–1380.
[169] Richard Isdahl and Odd Erik Gundersen. 2019. Out-of-the-box reproducibility: A survey of machine learning platforms. In *Proceedings of the 15th International Conference on Escience (eScience'19)*. IEEE, 86–95.
[170] Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3543–3556.
[171] Carolin E. M. Jakob, Florian Kohlmayer, Thierry Meurers, Jörg Janne Vehreschild, and Fabian Prasser. 2020. Design and evaluation of a data anonymization pipeline to promote open science on COVID-19. *Sci. Data* 7, 1 (2020), 1–10.
[172] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy artificial intelligence. *Govern. Inf. Quart.* 37, 3 (2020), 101493.
[173] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2020. Monitoring misuse for accountable'artificial intelligence as a service. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 300–306.
[174] Shouling Ji, Weiqing Li, Prateek Mittal, Xin Hu, and Raheem Beyah. 2015. {SecGraph}: A uniform and open-source evaluation system for graph data anonymization and de-anonymization. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security'15)*. 303–318.
[175] Shouling Ji, Weiqing Li, Mudhakar Srivatsa, and Raheem Beyah. 2014. Structural data de-anonymization: Quantification, practice, and implications. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1040–1053.
[176] Shouling Ji, Prateek Mittal, and Raheem Beyah. 2016. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Commun. Surv. Tutor.* 19, 2 (2016), 1305–1326.
[177] Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. 2014. Differential privacy and machine learning: A survey and review. arXiv:1412.7584. Retrieved from https://arxiv.org/abs/1412.7584.
[178] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 9 (2019), 389–399.
[179] Sushant Kafle, Abraham Glasser, Sedeeq Al-khazraji, Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Artificial intelligence fairness in the context of accessibility research on intelligent systems for people who are deaf or hard of hearing. *ACM SIGACCESS Access. Comput.* 125 (2020), 1–1.
[180] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* 14, 1–2 (2021), 1–210.
[181] Ehud Karnin, Jonathan Greene, and Martin Hellman. 1983. On secret sharing systems. *IEEE Trans. Inf. Theory* 29, 1 (1983), 35–41.

[182] 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*. Springer, 105–115.
[183] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2017. Generalization in deep learning. arXiv:1710.05468. Retrieved from https://arxiv.org/abs/1710.05468.
[184] Been Kim and F. Doshi-Velez. 2018. Introduction to interpretable machine learning. In *Proceedings of the CVPR'18 Tutorial on Interpretable Machine Learning for Computer Vision*.
[185] Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Adv. Neur. Inf. Process. Syst.* 29 (2016).
[186] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. *Adv. Neur. Inf. Process. Syst.* 31 (2018).
[187] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.
[188] Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. 2021. Crypten: Secure multi-party computation meets machine learning. *Adv. Neur. Inf. Process. Syst.* 34 (2021).
[189] Bran Knowles and John T. Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 262–271.
[190] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, Vol. 2.
[191] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U.S.A.* 117, 14 (2020), 7684–7689.
[192] Nathan Koenig and Andrew Howard. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vol. 3. IEEE, 2149–2154.
[193] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.
[194] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. 2020. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 301–310.
[195] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the World Wide Web Conference*. 853–862.
[196] Anders Krogh and John A. Hertz. 1992. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*. 950–957.
[197] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. 2020. Causal inference. *Engineering* 6, 3 (2020), 253–263.
[198] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.
[199] 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 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops'20)*. IEEE, 1–6.
[200] Nishant Kumar, Mayank Rathee, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. 2020. Cryptflow: Secure tensorflow inference. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'20)*. IEEE, 336–353.
[201] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2018. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*. Chapman & Hall/CRC, 99–112.
[202] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. In *The 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning*.
[203] Alexander Lavin, Ciarán M. Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, Atılım Güneş Baydin, Amit Sharma, Adam Gibson, et al. 2021. Technology readiness levels for machine learning systems. arXiv:2101.03989. Retrieved from https://arxiv.org/abs/2101.03989.
[204] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 656–672.
[205] Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, 64–69. https://doi.org/10.18653/v1/P19-3011

[206] Klas Leino, Zifan Wang, and Matt Fredrikson. 2021. Globally-robust neural networks. In *International Conference on Machine Learning*. PMLR, 6212–6222.

[207] Julio Cesar Sampaio do Prado Leite and Claudia Cappelli. 2010. Software transparency. *Bus. Inf. Syst. Eng.* 2, 3 (2010), 127–139.

[208] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philos. Technol.* 31, 4 (2018), 611–627.

[209] David Leslie. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute.

[210] Dave Lewis, David Filip, and Harshvardhan J. Pandit. 2021. An ontology for standardising trustworthy AI. In *Factoring Ethics in Technology, Policy Making, Regulation and AI*, 65.

[211] Bo Li, Yevgeniy Vorobeychik, and Xinyun Chen. 2016. A general retraining framework for scalable adversarial classification. In *NIPS 2016 Workshop on Adversarial Training*.

[212] Ke Li and Jitendra Malik. 2017. Learning to optimize. In *International Conference on Learning Representations*.

[213] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2017. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

[214] Yehuda Lindell. 2020. Secure multiparty computation. *Communications of the ACM* 64, 1 (2020), 86–96.

[215] Yehuda Lindell and Benny Pinkas. 2009. A proof of security of yao's protocol for two-party computation. *J. Cryptol.* 22, 2 (2009), 161–188.

[216] Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[217] Bo Liu, Chaowei Tan, Jiazhou Wang, Tao Zeng, Huasong Shan, Houpu Yao, Huang Heng, Peng Dai, Liefeng Bo, and Yanqing Chen. 2021. Fedlearn-algo: A flexible open-source privacy-preserving machine learning platform. arXiv:2107.04129. Retrieved from https://arxiv.org/abs/2107.04129.

[218] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Anil K. Jain, and Jiliang Tang. 2022. Trustworthy ai: A computational perspective. *ACM Trans. Intell. Syst. Technol*, Jun 2022. Just Accepted.

[219] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 273–294.

[220] Kun Liu and Evimaria Terzi. 2008. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 93–106.

[221] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. 2019. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3673–3682.

[222] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. 2020. A secure federated transfer learning framework. *IEEE Intell. Syst.* 35, 4 (2020), 70–82.

[223] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. 2020. Federated learning for open banking. In *Federated Learning*. Springer, 240–254.

[224] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[225] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Adv. Neur. Inf. Proc. Syst.* 30 (2017), 4765–4774.

[226] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, et al. 2018. Deepgauge: Multi-granularity testing criteria for deep learning systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 120–131.

[227] Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. 2021. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Comput. Surv.* 55, 1 (2021), 1–38.

[228] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

[229] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

[230] Silvia Magdici and Matthias Althoff. 2016. Fail-safe motion planning of autonomous vehicles. In *Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC'16)*. IEEE, 452–458.

[231] Pratyush Maini, Eric Wong, and Zico Kolter. 2020. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*. PMLR, 6640–6650.

[232] Abdul Majeed and Sungchang Lee. 2020. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access* 9 (2020), 8512–8545.

[233] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. 2021. Who needs mlops: What data scientists seek to accomplish and how can MLOps help? In *Proceedings of the IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN'21)*. IEEE, 109–112.
[234] Christopher D. Manning. 2022. Human language understanding & reasoning. *Daedalus* 151, 2 (2022), 127–138.
[235] Dusica Marijan and Arnaud Gotlieb. 2020. Software testing for machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13576–13582.
[236] David J. Martin, Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y. Halpern. 2007. Worst-case background knowledge for privacy-preserving data publishing. In *Proceedings of the IEEE 23rd International Conference on Data Engineering*. IEEE, 126–135.
[237] Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*. PMLR, 1614–1623.
[238] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.
[239] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5447–5456.
[240] Frank McKeen, Ilya Alexandrovich, Ittai Anati, Dror Caspi, Simon Johnson, Rebekah Leslie-Hurd, and Carlos Rozas. 2016. Intel software guard extensions (intel sgx) support for dynamic memory management inside an enclave. In *Proceedings of the Hardware and Architectural Support for Security and Privacy 2016*. 1–9.
[241] 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.
[242] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2021), 1–35.
[243] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. 2017. On detecting adversarial perturbations. In *5th International Conference on Learning Representations (ICLR'17), Toulon, France, April 24-26, 2017, Conference Track Proceedings*.
[244] Silvio Micali, Oded Goldreich, and Avi Wigderson. 1987. How to play any mental game. In *Proceedings of the 19th ACM Symposium on Theory of Computing (STOC'87)*. ACM, 218–229.
[245] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
[246] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.
[247] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 19–38.
[248] Sina Mohseni, Jeremy E. Block, and Eric D. Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. arXiv:1801.05075. Retrieved from https://arxiv.1801.05075.
[249] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interacti. Intell. Syst.* 11, 3-4 (2021), 1–45.
[250] Christoph Molnar. 2020. Interpretable Machine Learning. Retrieved from Lulu.com.
[251] Miranda Mourby, Elaine Mackey, Mark Elliot, Heather Gowans, Susan E. Wallace, Jessica Bell, Hannah Smith, Stergios Aidinlis, and Jane Kaye. 2018. Are 'pseudonymise' data always personal data? Implications of the GDPR for administrative data research in the UK. *Comput. Law Secur. Rev.* 34, 2 (2018), 222–233.
[252] Mark Wilfried Müller. 2016. *Increased Autonomy for Quadrocopter Systems: Trajectory Generation, Fail-safe Strategies and State Estimation*. Ph.D. Dissertation. ETH Zurich.
[253] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. 2007. Hiding the presence of individuals from shared databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 665–676.
[254] OpenAI. 2018. OpenAI Charter. Retrieved February 20, 2021 from https://openai.com/charter/.
[255] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2009), 1345–1359.
[256] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2020. Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training. arXiv:2007.02375. Retrieved from https://arxiv.org/abs/2007.02375.
[257] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Comput. Surv.* 54, 2 (2021), 1–38.
[258] Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. 2018. Label sanitization against label flipping poisoning attacks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 5–15.
[259] Judea Pearl. 2009. Causal inference in statistics: An overview. *Stat. Surv.* 3 (2009), 96–146.

[260] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *Proceedings of the 26th Symposium on Operating Systems Principles*. 1–18.

[261] Antonio Pérez, M. Isabel García, Manuel Nieto, José L. Pedraza, Santiago Rodríguez, and Juan Zamorano. 2010. Argos: An advanced in-vehicle data recorder on a massively sensorized vehicle for car driver behavior experimentation. *IEEE Trans. Intell. Transport. Syst.* 11, 2 (2010), 463–473.

[262] Zahid Pervaiz, Walid G. Aref, Arif Ghafoor, and Nagabhushana Prabhu. 2013. Accuracy-constrained privacy-preserving access control mechanism for relational data. *IEEE Trans. Knowl. Data Eng.* 26, 4 (2013), 795–807.

[263] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *J. Mach. Learn. Res.* 22 (2021).

[264] Sandro Pinto and Nuno Santos. 2019. Demystifying arm trustzone: A comprehensive survey. *ACM Comput. Surv.* 51, 6 (2019), 1–36.

[265] David Piorkowski, Daniel González, John Richards, and Stephanie Houde. 2020. Towards evaluating and eliciting high-quality documentation for intelligent systems. arXiv:2011.08774. Retrieved from https://arxiv.org/abs/2011.08774.

[266] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. *Adv. Neur. Inf. Process. Syst.* 30 (2017).

[267] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–52.

[268] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. *Adv. Neur. Inf. Process. Syst.* 32 (2019).

[269] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[270] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

[271] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. In *ICML 2019 Workshop Deep Phenomena*.

[272] 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 Conference on Fairness, Accountability, and Transparency*. 33–44.

[273] Inioluwa Deborah Raji and Jingying Yang. 2019. About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. In *Human-Centric Machine Learning workshop at Neural Information Processing Systems conference 2019*.

[274] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (2021), 1–23.

[275] Raghavendra Ramachandra and Christoph Busch. 2017. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.* 50, 1 (2017), 1–37.

[276] Deevashwer Rathee, Mayank Rathee, Nishant Kumar, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. 2020. CrypTFlow2: Practical 2-party secure inference. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 325–342.

[277] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. Retrieved from https://ainowinstitute.org/aiareport2018.pdf.

[278] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A McCann, and S. Yu Philip. 2018. LoPub: High-dimensional crowdsourced data publication with local differential privacy. *IEEE Trans. Inf. Forens. Secur.* 13, 9 (2018), 2151–2166.

[279] 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.

[280] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[281] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4902–4912.

[282] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. 2020. The future of digital health with federated learning. *NPJ Digit. Med.* 3, 1 (2020), 1–7.

[283] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* 10, 1 (2019), 1–9.

[284] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2020. FR-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*. PMLR, 8147–8157.

[285] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Auton. Agents Multi-Agent Syst.* 33, 6 (2019), 673–705.

[286] Andrew Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[287] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: What it is, and what it is not. In *Proceedings of the IEEE International Conference on Trust, Security, and Privacy in Computing and Communications, IEEE International Conference on Big Data Science and Engineering, and IEEE International Symposium on Parallel and Distributed Processing with Applications (Trustcom/BigDataSE/ISPA'15)*, Vol. 1. IEEE, 57–64.

[288] Pierangela Samarati and Latanya Sweeney. 1998. Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

[289] Elena Samuylova. 2020. Machine Learning Monitoring: What It Is, and What We Are Missing. Retrieved May 18, 2021 from https://towardsdatascience.com/machine-learning-monitoring-what-it-is-and-what-we-are-missing-e644268023ba.

[290] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of the Data and Discrimination: Converting Critical Concerns Into Productive Inquiry* 22 (2014), 4349–4357.

[291] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*. PMLR, 1842–1850.

[292] Laura Schelenz, Avi Segal, and Kobi Gal. 2020. Applying transparency in artificial intelligence based personalization systems. arXiv:2004.00935. Retrieved from https://arxiv.org/abs/2004.00935.

[293] Sebastian Schelter, Joos-Hendrik Boese, Johannes Kirschnick, Thoralf Klein, and Stephan Seufert. 2017. Automatically tracking metadata and provenance of machine learning experiments. In *Machine Learning Systems Workshop at NIPS*. 27–29.

[294] Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas. 2021. AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Trans. Technol. Soc.* (2021).

[295] Albrecht Schmidt and Thomas Herrmann. 2017. Intervention user interfaces: A new interaction paradigm for automated systems. *Interactions* 24, 5 (2017), 40–45.

[296] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Adv. Neur. Inf. Process. Syst.* 28 (2015), 2503–2511.

[297] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[298] Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. 2022. Toward verified artificial intelligence. *Communications of the ACM* 65, 7 (2022), 46–55.

[299] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, et al. 2020. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* 10, 1 (2020), 1–12.

[300] Haoyi Shi, Chao Jiang, Wenrui Dai, Xiaoqian Jiang, Yuzhe Tang, Lucila Ohno-Machado, and Shuang Wang. 2016. Secure multi-pArty computation grid LOgistic REgression (SMAC-GLORE). *BMC Med. Inf. Decis. Mak.* 16, 3 (2016), 175–187.

[301] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Trans. Interact. Intell. Syst.* 10, 4 (2020), 1–31.

[302] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*. PMLR, 3145–3153.

[303] Umair Siddique. 2020. SafetyOps. arXiv:2008.04461. Retrieved from https://arxiv.org/abs/2008.04461.

[304] Samuel Henrique Silva and Peyman Najafirad. 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. arXiv:2007.00753. Retrieved from https://arxiv.org/abs/2007.00753.

[305] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR'14)*.

[306] Jatinder Singh, Jennifer Cobbe, and Chris Norval. 2018. Decision provenance: Harnessing data flow for accountable systems. *IEEE Access* 7 (2018), 6562–6574.
[307] Tomáš Sixta, Julio Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. 2020. Fairface challenge at eccv 2020: Analyzing bias in face recognition. In *European Conference on Computer Vision*. Springer, 463–481.
[308] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Adv. Neur. Inf. Process. Syst.* 30 (2017).
[309] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.
[310] Brian Stanton, Theodore Jensen, et al. Trust and artificial intelligence. 2021. Draft NISTIR 8332, National Institute: of Standards and Technology.
[311] Simone Stumpf, Adrian Bussone, and Dympna O'sullivan. 2016. Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.
[312] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2018. Is robustness the cost of accuracy?—A comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 631–648.
[313] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *Proceedings of the International Conference on Learning Representations*.
[314] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.
[315] Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. 2020. Auditing Machine Learning Algorithms. Retrieved from https://www.auditingalgorithms.net/index.html.
[316] Ki Hyun Tae, Yuji Roh, Young Hun Oh, Hyunsu Kim, and Steven Euijong Whang. 2019. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*. 1–4.
[317] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. 2020. Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the ACM-IMS on Foundations of Data Science Conference*. 23–34.
[318] Xin Tan, Yiheng Zhang, Ying Cao, Lizhuang Ma, and Rynson W. H. Lau. 2021. Night-time sceneparsing with a large real dataset. *IEEE Transactions on Image Processing*, 30 (2021), 9085–9098.
[319] Brad Templaton. 2019. Tesla's "Shadow" Testing Offers a Useful Advantage on the Biggest Problem in Robocars. Retrieved June 15, 2021 from https://www.forbes.com/sites/bradtempleton/2019/04/29/teslas-shadow-testing-offers-a-useful-advantage-on-the-biggest-problem-in-robocars/?sh=7a960b9e3c06.
[320] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. 2008. Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.* 1, 1 (2008), 115–125.
[321] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. 2010. Cost-sensitive learning methods for imbalanced data. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'10)*. IEEE, 1–8.
[322] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2020. Trustworthy artificial intelligence. *Electr. Markets* (2020), 1–18.
[323] Martijn Tideman. 2010. Scenario-based simulation environment for assistance systems. *ATZautotechnology* 10, 1 (2010), 28–32.
[324] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5026–5033.
[325] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. 2019. Improving robustness of {ML} classifiers against realizable evasion attacks using conserved features. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security'19)*. 285–302.
[326] Cecília Torres, Walter Franklin, and Laura Martins. 2018. Accessibility in chatbots: The state of the art in favor of users with visual impairment. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 623–635.
[327] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P'17)*. IEEE, 401–416.
[328] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *Proceedings of the International Conference on Learning Representations*.
[329] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security'16)*. 601–618.
[330] Amos Treiber, Alejandro Molina, Christian Weinert, Thomas Schneider, and Kristian Kersting. 2020. Cryptospn: Privacypreserving sum-product network inference. In *24th European Conference on Artificial Intelligence (ECAI) 2020, Santiago de Compostela, Spain*.

[331] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Mądry. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations (ICLR'19), New Orleans, LA, USA, May 6-9, 2019*.

[332] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9073–9080.

[333] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics Inf. Technol.* 11, 2 (2009), 105–112.

[334] UNI Global Union. 2017. Top 10 principles for ethical artificial intelligence. In *The Future World of Work*.

[335] Caterina Urban and Antoine Miné. 2021. A review of formal methods applied to machine learning. arXiv:2104.02466. Retrieved from https://arxiv.org/abs/2104.02466.

[336] Joaquin Vanschoren. 2018. Meta-learning: A survey. arXiv:1810.03548. Retrieved from https://arxiv.org/abs/1810.03548.

[337] Vladimir Vapnik. 2013. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

[338] Kush R. Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads* 25, 3 (2019), 26–29.

[339] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (Fairware'18)*. IEEE, 1–7.

[340] Yevgeniy Vorobeychik and Murat Kantarcioglu. 2018. Adversarial machine learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 12, 3 (2018), 1–169.

[341] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. J. Law Technol.* 31 (2017), 841.

[342] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 707–723.

[343] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. 2021. Generalizing to unseen domains: A survey on domain generalization (unpublished).

[344] Qian Wang, Yan Zhang, Xiao Lu, Zhibo Wang, Zhan Qin, and Kui Ren. 2016. Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. *IEEE Trans. Depend. Sec. Comput.* 15, 4 (2016), 591–606.

[345] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5329–5336.

[346] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. 2019. On the convergence and robustness of adversarial training. In *Proceedings of the International Conference on Machine Learning (ICML'19)*, Vol. 1. 2.

[347] Yingxu Wang and Jingqiu Shao. 2003. Measurement of the cognitive functional complexity of software. In *Proceedings of the 2nd IEEE International Conference on Cognitive Informatics*. IEEE, 67–74.

[348] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* 53, 3 (2020), 1–34.

[349] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forens. Secur.* 15 (2020), 3454–3469.

[350] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *J. Big Data* 3, 1 (2016), 1–40.

[351] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[352] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the robustness of neural networks: An extreme value theory approach. In *Proceedings of the International Conference on Learning Representations*.

[353] Chathurika S. Wickramasinghe, Daniel L. Marino, Javier Grandio, and Milos Manic. 2020. Trustworthy AI development guidelines for human system interaction. In *Proceedings of the 13th International Conference on Human System Interaction (HSI'20)*. IEEE, 130–136.

[354] Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[355] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 11–20.

[356] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[357] Jeannette M. Wing. 2021. Trustworthy ai. *Commun. ACM* 64, 10 (2021), 64–71.

[358] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. 2020. Defending against physically realizable attacks on image classification. In *8th International Conference on Learning Representations (ICLR'20), Addis Ababa, Ethiopia, April 26-30, 2020*.

[359] Bernhard Wymann, Eric Espié, Christophe Guionneau, Christos Dimitrakakis, Rémi Coulom, and Andrew Sumner. 2000. Torcs, the open racing car simulator. Retrieved from http://torcs.sourceforge.net.

[360] Weiyi Xia, Yongtai Liu, Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantacioglu, Steve Nyemba, Ellen Wright Clayton, and Bradley A. Malin. 2021. Enabling realistic health data re-identification risk assessment through adversarial modeling. *J. Am. Med. Inf. Assoc.* 28, 4 (2021), 744–752.

[361] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*. PMLR, 11492–11501.

[362] Huan Xu and Shie Mannor. 2012. Robustness and generalization. *Mach. Learn.* 86, 3 (2012), 391–423.

[363] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. PMLR, 2048–2057.

[364] Qian Xu, Md Tanvir Arafin, and Gang Qu. 2021. Security of neural networks from hardware perspective: A survey and beyond. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC'21)*. IEEE, 449–454.

[365] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* 10, 2 (2019), 1–19.

[366] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2369–2380.

[367] Andrew C. Yao. 1982. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS'82)*. IEEE, 160–164.

[368] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. 2021. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, Vol. 2. 7.

[369] Yu Yao and Ella Atkins. 2020. The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Trans. Intell. Transport. Sys.* (2020).

[370] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constr. Approx.* 26, 2 (2007), 289–315.

[371] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT'18)*.

[372] Mingxuan Yuan, Lei Chen, S. Yu Philip, and Ting Yu. 2011. Protecting sensitive labels in social network data anonymization. *IEEE Trans. Knowl. Data Eng.* 25, 3 (2011), 633–647.

[373] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 9 (2019), 2805–2824.

[374] Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, et al. 2018. Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.* 41, 4 (2018), 39–45.

[375] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. Springer, 818–833.

[376] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. PMLR, 325–333.

[377] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations (ICLR'17), Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[378] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, and Cho Jui Hsieh. 2019. Towards stable and efficient training of verifiably robust neural networks. *J. Environ. Sci. (Chin.)* (2019).

[379] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. 2018. Efficient neural network robustness certification with general activation functions. *Adv. Neural Inf. Process. Syst.* 31 (2018).
[380] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*. PMLR, 7472–7482.
[381] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2022. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 1 (2022), 1–36.
[382] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. 2018. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In *Proceedings of the 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE'18)*. IEEE, 132–142.
[383] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8827–8836.
[384] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6261–6270.
[385] Sicong Zhang, Hui Yang, and Lisa Singh. 2016. Anonymizing query logs by differential privacy. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 753–756.
[386] Xuyun Zhang, Laurence T. Yang, Chang Liu, and Jinjun Chen. 2013. A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. *IEEE Trans. Parallel Distrib. Syst.* 25, 2 (2013), 363–373.
[387] Chuan Zhao, Shengnan Zhao, Minghao Zhao, Zhenxiang Chen, Chong-Zhi Gao, Hongwei Li, and Yu-an Tan. 2019. Secure multi-party computation: Theory, practice and applications. *Inf. Sci.* 476 (2019), 357–372.
[388] Ziyuan Zhong. A Tutorial on Fairness in Machine Learning. Retrieved from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.
[389] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.
[390] Bin Zhou, Jian Pei, and WoShun Luk. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explor. Newslett.* 10, 2 (2008), 12–22.
[391] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
[392] Yichen Zhou and Giles Hooker. 2016. Interpreting models via single tree approximation. arXiv:1610.09036. Retrieved from https://arxiv.org/abs/1610.09036.
[393] Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. 2003. Extracting symbolic rules from trained neural network ensembles. *Ai Commun.* 16, 1 (2003), 3–15.
[394] Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama. 2016. An overview of concept drift applications. *Big Data Anal.: New Algor. New Soc.* (2016), 91–114.

Received 3 October 2021; revised 6 May 2022; accepted 2 August 2022