Trustworthy AI: A Computational Perspective

HAOCHEN LIU*, Michigan State University, USA
YIQI WANG*, Michigan State University, USA
WENQI FAN, The Hong Kong Polytechnic University, Hong Kong
XIAORUI LIU, Michigan State University, USA
YAXIN LI, Michigan State University, USA
SHAILI JAIN, Twitter, USA
YUNHAO LIU, Tsinghua University, China
ANIL K. JAIN, Michigan State University, USA
JILIANG TANG, Michigan State University, USA

In the past few decades, artificial intelligence (AI) technology has experienced swift developments, changing everyone's daily life and profoundly altering the course of human society. The intention behind developing AI was and is to benefit humans by reducing labor, increasing everyday conveniences, and promoting social good. However, recent research and AI applications indicate that AI can cause unintentional harm to humans by, for example, making unreliable decisions in safety-critical scenarios or undermining fairness by inadvertently discriminating against a group or groups. Consequently, trustworthy AI has recently garnered increased attention regarding the need to avoid the adverse effects that AI could bring to people, so people can fully trust and live in harmony with AI technologies.

A tremendous amount of research on trustworthy AI has been conducted and witnessed in recent years. In this survey, we present a comprehensive appraisal of trustworthy AI from a computational perspective to help readers understand the latest technologies for achieving trustworthy AI. Trustworthy AI is a large and complex subject, involving various dimensions. In this work, we focus on six of the most crucial dimensions in achieving trustworthy AI: (i) Safety & Robustness, (ii) Nondiscrimination & Fairness, (iii) Explainability, (iv) Privacy, (v) Accountability & Auditability, and (vi) Environmental Well-being. For each dimension, we review the recent related technologies according to a taxonomy and summarize their applications in real-world systems. We also discuss the accordant and conflicting interactions among different dimensions and discuss potential aspects for trustworthy AI to investigate in the future.

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

Additional Key Words and Phrases: artificial intelligence, robustness, fairness, explainability, privacy, accountability, environmental well-being

*Both authors contributed equally to this work.

Manuscript submitted to ACM

ACM Reference Format:
Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil K. Jain, and Jiliang Tang. 2018. Trustworthy AI: A Computational Perspective. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 57 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Artificial intelligence (AI), a science that studies and develops the theory, methodology, technology, and application systems for simulating, extending, and expanding human intelligence, has brought revolutionary impact to modern human society. From a micro view, AI plays an irreplaceable role in many aspects of our lives. Modern life is filled with interactions with AI applications: from unlocking a cell phone with a face ID, talking to a voice assistant to buying products recommended by e-commerce platforms; from a macro view, AI creates great economic outcomes. The Future of Jobs Report 2020 from the World Economic Forum [144] predicts that AI will create 58 million new jobs in five years. By 2030, AI is expected to produce extra economic profits of 13 trillion U.S. dollars, which contribute 1.2% annual growth to the GDP of the whole world [58]. However, along with their rapid and impressive development, AI systems have also exposed their untrustworthy sides. For example, safety-critical AI systems are shown to be vulnerable to adversarial attacks. Deep image recognition systems in autonomous vehicles could fail to distinguish road signs modified by malicious attackers [364], posing a great threat to passenger safety. In addition, AI algorithms can cause bias and unfairness. Online AI chatbots could produce indecent, racist, and sexist content [354] that offends users and has a negative social impact. Moreover, AI systems carry the risk of disclosing user privacy and business secrets. Hackers can take advantage of the feature vectors produced by an AI model to reconstruct private input data, such as fingerprints [26], thereby leaking a user's sensitive information. These vulnerabilities can make existing AI systems unusable and can cause severe economic and security consequences. Concerns around trustworthiness have become a huge obstacle to the boosting of AI usage and increasing of economic value. Hence, how to build trustworthy AI systems has become a focal topic in both academia and industry.

In recent years, a large body of literature on trustworthy AI has emerged. With the increasing demand for building trustworthy AI, it is imperative to summarize existing achievements and discuss possible directions for future research. In this survey, we provide a comprehensive overview of trustworthy AI to help newcomers attain a basic understanding of what makes an AI system trustworthy and to help veterans track the latest progress in the field. We clarify the definition of trustworthy AI and introduce six key dimensions of it. For each dimension, we present its concepts and taxonomies and review representative algorithms. We also introduce possible interactions among different dimensions and discuss other potential issues around trustworthy AI that have not yet drawn sufficient attention. In addition to definitions and concepts, our survey focuses on the specific computational solutions for realizing each dimension of trustworthy AI. This perspective makes it distinct from some extant related works, such as a government guideline [325], which suggests how to build a trustworthy AI system in the form of laws and regulations, or reviews [55, 337], which discuss the realization of trustworthy AI from a high-level, non-technical perspective.

According to a recent ethics guideline for AI provided by the European Union (EU) [325], a trustworthy AI system should meet four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. Based on these four principles, AI researchers, practitioners, and governments propose various specific dimensions for trustworthy AI [55, 325, 337]. In this survey, we focus on six important and concerning dimensions that have been extensively studied. As shown in Figure 1, they are **Safety & Robustness, Non-discrimination & Fairness, Explainability, Privacy, Auditability & Accountability, and Environmental Well-Being**. These dimensions are

Fig. 1. Six key dimensions of trustworthy AI.

very vital in real applications. Take the face recognition system as an example, it has been widely adopted in identity authentication. Thus, the face recognition system is desired to make highly robust and accurate predictions under any potential risks and attacks. Meanwhile, human face information is very private and crucial biometric information. It should be carefully protected. Furthermore, it is important to eliminate demographic disparities existent in the performance of face recognition, so that the system shows no bias towards any certain ethnic groups or genders. Also, for the reliability of the face recognition system, it is preferred that the recognition result can be explained in a reasonable way and the system can be audited periodically. In addition, the training process of large-scale face recognition models and the wide deployment of them require huge energy consumption, which may lead to a great amount of carbon emission. Thus, how to ensure the environmental friendliness of face recognition systems is important.

The remaining survey is organized as follows. In section 2, we articulate the definition of trustworthy AI and provide various definitions of it to help readers understand how a trustworthy AI system is defined by researchers from such different disciplines as computer science, sociology, law, and business. We then distinguish the concept of trustworthy AI from several related concepts, such as ethical AI and responsible AI.

In section 3, we detail the dimension of **Safety & Robustness**, which requires an AI system to be robust to the tiny perturbations of inputs and to be able to make secure decisions. In recent years, numerous studies have shown that AI systems, especially those that adopt deep learning models, can be very sensitive to intentional or unintentional inputs perturbations, posing huge risks to safety-critical applications. For example, as described before, autonomous vehicles can be fooled by altered road signs. Additionally, spam detection models can be fooled by emails with well-designed text [32]. Thus, spam senders can take advantage of this weakness to make their emails immune to the detection system, which would cause a bad user experience.

It has been demonstrated that AI algorithms can learn human discrimination through provided training examples and make unfair decisions. For example, some face recognition algorithms have difficulty detecting faces of African Americans [297] or misclassifying them as gorillas [178]. Moreover, voice dictation software typically performs better

at recognizing a voice from a male than that from a female [294]. In section 4, we introduce the dimension of **Non-discrimination & Fairness** in which an AI system is expected to avoid unfair bias toward certain groups or individuals.

In section 5, we discuss the dimension of **Explainability**, which suggests that the AI's decision mechanism system should be able to be explained to stakeholders (who should be able to understand the explanation). For example, AI techniques have been used for disease diagnosis based on the symptoms and physical features of a patient [306]. In such cases, a black-box decision is not acceptable. The inference process should be transparent to doctors and patients to ensure that the diagnosis is exact in every detail.

Researchers have found that some AI algorithms can store and expose users' personal information. For example, dialogue models trained on the human conversation corpus can remember sensitive information, like credit card numbers, which can be elicited by interacting with the model [173]. In section 6, we present the dimension of **Privacy**, which requires an AI system to avoid leaking any private information.

In section 7, we describe the dimension of **Auditability & Accountability**, which expects that an AI system is assessed by a third party and, when necessary, assign responsibility for an AI failure, especially in critical applications [325].

Recently, the environmental impacts of AI systems have drawn people's attention, since some large AI systems consume great amounts of energy. As a mainstream AI technology, deep learning is moving toward pursuing larger models and more parameters. Accordingly, more storage and computational resources are consumed. A study [330] shows that training a BERT model [116] costs a carbon emission of around 1,400 pounds of carbon dioxide, which is comparable to that of a round trip trans-America flight. Therefore, an AI system should be sustainable and environmentally friendly. In section 8, we review the dimension of **Environmental Well-Being**.

In section 9, we discuss the interactions among the different dimensions. Recent studies have demonstrated that there are accordance and conflicts among different dimensions of trustworthy AI [325, 352]. For example, the robustness and explainability of deep neural networks are tightly connected and robust models tend to be more interpretable [128, 342] and vice versa [269]. Moreover, it is shown that in some cases, a trade-off exists between robustness and privacy. For instance, adversarial defense approaches can make a model more vulnerable to membership inference attacks, which increases the risk of training data leakage [326].

In addition to the aforementioned six dimensions, there are more dimensions of trustworthy AI, such as human agency and oversight, creditability, etc. Although these additional dimensions are as important as the six dimensions considered in this article, they are in earlier stages of development with limited literature, especially for computational methods. Thus, in section 10, we discuss these dimensions of trustworthy AI as future directions needing dedicated research efforts.

2 CONCEPTS AND DEFINITIONS

The word "trustworthy" is noted to mean "worthy of trust of confidence; reliable, dependable" in the Oxford English Dictionary or "able to be trusted" in the Dictionary of Cambridge. "Trustworthy" descends from the word trust, which is described as the "firm belief in the reliability, truth, or ability of someone or something" in the Oxford English Dictionary or the "belief that you can depend on someone or something" in the Dictionary of Cambridge. Broadly speaking, trust is a widespread notion in human society, which lays the important foundation for the sustainable development of human civilization. Strictly speaking, some potential risks always exist in our external environment because we cannot completely control people and other entities [248, 337]. It is our trust in these parties that allows us to put ourselves at potential risk to continue interacting with them willingly [219]. Trust is necessary among people. It forms the basis of a

good relationship and is necessary for people to live happily and to work efficiently together. In addition, trust is also vital between humans and technology. Without trust, humans would not be willing to utilize technology, which would undoubtedly impede its advancement and prevent people from enjoying the conveniences it brings. Therefore, for a win-win situation between humans and technology, it is necessary to guarantee that the technology is trustworthy so people can build trust in it.

The term "artificial intelligence" got its name from a workshop at a 1956 Dartmouth conference [57, 249]. Although there are numerous definitions for AI [210], AI generally denotes a program or system that is able to cope with a real-world problem with human-like reasoning, for example, in the field of image recognition within AI, which uses deep learning networks to recognize objects or people within images [288]. The past few decades have witnessed rapid and impressive development of AI; there are tons of breakthroughs happening in every corner of this field [301, 332]. Furthermore, with the rapid development of big data and computational resources, AI has been broadly applied to many aspects of human life, including economics, healthcare, education, transportation, and so on, where it has revolutionized industries and achieved numerous feats. Considering the important role AI plays in modern society, it is necessary to make AI trustworthy so that humans can rely on it with minimal concern regarding its potential harm. Trust is essential in allowing the potential of AI to be fully realized – and humans to fully enjoy its benefits and convenience [98].

Table 1. A summary of principles for Trustworthy AI from different perspectives.

Perspective	Principles
Technical	Accuracy, Robustness, Explainability
User	Availability, Usability, Safety, Privacy, Autonomy
Social	Law-abiding, Ethical, Fair, Accountable, Environmental-friendly

Due to its importance and necessity, trustworthy AI has drawn increasing attention, and there are numerous discussions and debates over its definition and extension [98]. In this survey, we define trustworthy AI as *programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm.* We further define trustworthy AI from the following three perspectives: the technical perspective, the user perspective, and the social perspective. An overall description of these perspectives is summarized in Table 1.

- **From a technical perspective**, trustworthy AI is expected to show the properties of accuracy, robustness, and explainability. Specifically, AI programs or systems should generate accurate output consistent with the ground truth as much as possible. This is also the first and most basic motivation for building them. Additionally, AI programs or systems should be robust to changes so that perturbations would not affect the model outcome. This is very important, since real environments where AI systems are deployed are usually very complex and volatile. Last, but not least, trustworthy AI must allow for explanation and analysis by humans, so that potential risks and harm can be minimized. In addition, trustworthy AI should be transparent so people can better understand its mechanism.
- **From a user's perspective**, trustworthy AI should possess the properties of availability, usability, safety, privacy, and autonomy. Specifically, AI programs or systems should be available for people whenever they need them, and these programs or systems should be easy to use for people with different backgrounds. More importantly, AI programs or systems are expected to avoid harm under any conditions, and to always put the safety of users

as the first priority. In addition, trustworthy AI should protect the privacy of all users. It should deal with data storage very carefully and seriously. Last but not least, the autonomy of trustworthy AI should always be under people's control. In other words, it is always a human's right to grant an AI system any decision-making power or to withdraw that power at any time.

- **From a social perspective**, trustworthy AI should be law-abiding, ethical, fair, accountable, and environmentally friendly. Specifically, AI programs or systems should operate in full compliance with all relevant laws and regulations and comply with the ethical principles of human society. Importantly, trustworthy AI should show nondiscrimination toward people from various backgrounds. It should guarantee justice and fairness among all users. Also, trustworthy AI should be accountable, which means it is clear who is responsible for each part of the AI system. Lastly, for the sustainable development and long-term prosperity of our civilization, AI programs and systems should be environmentally friendly. For example, they should limit energy consumption and cause minimal pollution.

Note that the above properties of the three perspectives are not independent of each other. Instead, they complement and reinforce each other.

There have been numerous terminologies related to AI proposed recently, including ethical AI, beneficial AI, responsible AI, explainable AI, fair AI, and so on. These terminologies share some overlap and distinction with trustworthy AI in terms of the intention and extension of the concept.

Next, we briefly describe some related terminologies to help readers enhance the understanding of trustworthy AI.

- **Ethical AI** [143]: An ethical framework of AI that specifies five core principles, including beneficence, nonmaleficence, autonomy, justice, and explicability. Additionally, 20 specific action points from four categories have been proposed to ensure continuous and effective efforts. They are assessment, development, incentivization, and support.
- **Beneficial AI** [271]: AI has undoubtedly brought people countless benefits, but to gain sustainable benefits from AI, 23 principles have been proposed in conjunction with the 2017 Asilomar conference. These principles are based on three aspects: research issues, ethics and values, and longer-term issues.
- **Responsible AI** [2, 4]: A framework for the development of responsible AI consists of 10 ethical principles: well-being, respect for autonomy, privacy and intimacy, solidarity, democratic participation, equity, diversity inclusion, prudence, responsibility, and sustainable development. The Chinese National Governance Committee for the New Generation Artificial Intelligence has proposed a set of governance principles to promote the healthy and sustainable development of responsible AI. Eight principles have been listed as follows: harmony and human-friendliness, fairness and justice, inclusion and sharing, respect for privacy, safety and controllability, shared responsibility, open and collaborative, and agile governance.
- **Explainable AI** [18]: The basic aim of explainable AI is to open up the "black box" of AI, to offer a trustworthy explanation of AI to users. It also aims to propose more explainable AI models, which can provide promising model performance and can be explained in non-technical terms at the same time, so that users can fully trust them and take full advantage of them.
- **Fair AI** [403]: Because AI is designed by humans and data plays a key role in most AI models, it is easy for AI to inherit some bias from its creators or input data. Without proper guidance and regulations, AI could be biased and unfair toward a certain group or groups of people. Fair AI denotes AI that shows no discrimination toward

people from any group. Its output should have little correlation with the traits of individuals, such as gender and ethnicity.

Overall, trustworthy AI has a very rich connotation and can be interpreted from several perspectives. It contains the concepts of many existing terminologies, including fair AI [403], explainable AI [18], and so on. Huge overlaps also exist among the concept of trustworthy AI and the concepts of ethical AI [143], beneficial AI [271], and responsible AI [2, 4]. Although proposed by different organizations, all of them aim at building reliable AI that sustainably benefits human society and these concepts can be exchangeable in some contexts. Note that Ethical AI [143] was proposed based on six documents including those where Beneficial AI [271] and Responsible AI [2] were proposed. Nonetheless, some differences exist among these concepts. For example, in beneficial AI and responsible AI, there are some principles and requirements for governments around the world, such as to avoid any arms race, but other concepts do not involve them.

3 SAFETY & ROBUSTNESS

A human can trust AI systems in safety-critical scenarios only if the machine learning (ML) systems can achieve stable and sustained high accuracy under the worst circumstances without any vulnerability. Otherwise, ML systems will cause severe consequences in safety-critical scenarios. To be specific, payment apps need to verify user identities through the face recognition system, but the identity theft may create different face images to fool the system and commit the crime [84]; ML systems are also widely used in developing autonomous driving techniques. Dangers may happen when the traffic sign recognition process generates a wrong prediction on the stop sign being slightly perturbed by some white mark [130]. In social networks, someone may pretend to be an important user by connecting with selected people to fool the social media mining algorithms. To avoid such dangerous behaviors, the ML systems should be robust to small perturbations since real-world data contains diverse types of noise. In recent years, many studies have shown that ML models can be fooled by imperceptible perturbations, namely, adversarial perturbations [243, 333]. From traditional ML classifiers [40] to deep learning models, like CNN [333], GNN [311], or RNN [381], none of the models is sufficiently robust to such perturbations. This raises huge concerns when ML models are applied to safety-critical tasks, such as authentication [84], autonomous driving [324], recommendation [132, 137], AI healthcare [142], etc. To build safe and reliable ML models, studying adversarial examples and the underlying reasons is urgent and essential.

In this section, we aim to introduce the concept of robustness, including how to design threat models and different types of defense strategies. We first introduce the concept of adversarial robustness. Then, we provide more details by introducing the taxonomy as well as examples for each category. We then discuss different adversarial attacks and defense strategies by introducing some representative methods. Next, we introduce how adversarial robustness issues affect real-world AI systems. We also present related tools and surveys to help readers further explore this field. Finally, we demonstrate potential future directions in adversarial robustness.

3.1 Concepts and Taxonomy

In this subsection, we describe the common and fundamental concepts and taxonomy in AI robustness to illustrate an overview of adversarial attacks and defenses.

3.1.1 Threat Models. An adversarial threat model is an adversarial attacker that tries to break the performance of ML models with fake training or test examples. The existence of adversarial attacks could lead to serious security concerns in a wide range of ML applications. Attackers use many different types of strategies to achieve their goals.

Therefore, threat models can be categorized into different types. In this subsection, we introduce different categories of threat models aiming to attack the ML system from different perspectives, including when the attack happens, what information of the model the attacker can access, and what the adversary's goal is.

- **Poisoning Attacks vs. Evasion Attacks.** Whether an attack is evasion or poisoning depends on whether attackers modify the training or test samples. A *poisoning attack* occurs when attackers add fake samples into the training set of a ML model. These fake samples are designed intentionally to train a bad model, thus the model would achieve bad overall performance [41] or give wrong predictions on a certain group of test samples [404]. When an adversary has access to the training data, this type of attack can happen and will raise realistic safety concerns. For example, the training data for an online shopping recommendation system is often collected from web users where attackers may exist. A special case of poisoning attacks is backdoor attack, which aims to only mislead the performance of test samples with a special trigger known by the attacker [84]. Recently, the connection between poisoning attack and shortcut learning[149, 174, 376] has been noticed. Shortcut learning summarizes a general phenomenon that machine learning models tend to link spurious features with the predictions in the training data while they could not generalize well on the test set. This aligns well with the underlying reason why poisoning attack leads to bad test behaviors.
  An *evasion attack* happens in the test phase. Given a well-trained classifier, attackers aim to design small perturbations for test samples in order to elicit wrong predictions from a victim model. From Figure 2, we can see that the image of a panda can be correctly classified by the model, while the perturbed version will be classified as a gibbon.

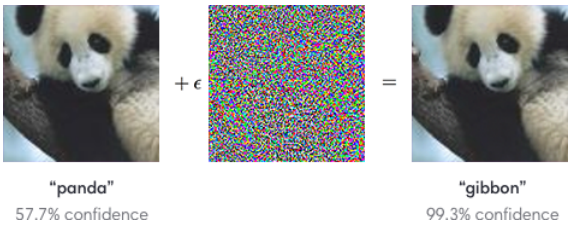


Fig. 2. An example of evasion attack. (Image Credit: [160])

- **White-box attacks vs. Black-box attacks.** According to the adversary's knowledge, attacking methods can be categorized into white-box and black-box attacks. *White-Box attacks* refer to a setting in which the adversary can utilize all the information of the target model, including the architecture, parameters, gradient information, etc. Generally speaking, the attacking process can be formulated as an optimization problem to optimize the risk of wrong predictions [69, 160]. With the ability to access white-box information, this problem is often much easier to be solved via gradient-based methods. White-box attacks have been extensively studied because the disclosure of model architecture and parameters helps people understand the weakness of ML models clearly; thus, they can be analyzed mathematically.
  In the *black-Box attack* setting, no knowledge of ML models is available to the adversaries. Adversaries can only feed the input data and query the outputs of the models. One of the most common ways to perform black-box attacks is to keep querying the victim model and approximate the gradient through numerical differentiation methods. Compared to white-box attacks, black-box attacks are more practical because ML models are less likely to be white-box due to privacy issues in reality.

- **Targeted Attacks vs. Non-Targeted Attacks.** In the image classification problem, threat models can be categorized by whether the adversary wants to get a pre-set label for a certain image. In *targeted attacks*, a specified target prediction label is expected for each adversarial example in the test phase. For example, identity theft may want to fool the face recognition system and pretend to be a specific important user.
  In contrast, *non-targeted attacks* expect an arbitrary prediction label except for the real one. Note that in other data domains, like graph data, the definition of targeted attack can be extended to mislead certain groups of nodes, but not necessary to force the model to give a particular prediction for one node.

3.1.2 Victim Models. Victim models are the models that are attacked by the attacker. The victim model ranges from traditional machine learning models like SVM [41] to Deep Neural Networks (DNNs), including the Convolution Neural Network (CNN) [217], Graph Neural Network (GNN) [311], Recurrent Neural Network (RNN) [381], etc. In this section, we briefly introduce the victim models that have been studied and shown to be vulnerable to adversarial attacks.

- **Traditional machine learning models.** One of the earliest works about robustness checked the security of Naive Bayes classifiers [109]. Later, SVM and the naive fully-connected neural networks have been shown to be vulnerable to attacks [40]. Recently, the adversarial robustness of tree-based models has also been proposed as an open problem [78].
- **Deep learning models.** In computer vision tasks, Convolution Neural Networks (CNNs) [212] are one of the most widely used models for image classification problems. CNN models aggregate local features from images to learn the representations of image objects and give predictions based on learned representations. The vulnerability of deep neural networks to attack was first proposed in CNN [333]. Since then, there have been extensive works indicating that CNNs are not robust against adversarial attacks. As another type of deep learning model, Graph Neural Networks (GNNs) have been developed for graph-structured data and can be used by many real-world systems, such as social networks and natural science. People also pay effort into testing GNN's robustness [45, 81, 108, 242, 406] and building robust GNNs [190]. Consider the node classification problem as an example: existing works show that the performance can be reduced significantly by slightly modifying node features, adding or deleting edges, or adding fake nodes [404]. For sequence models that deal with text data, there exist studies focusing on evaluating their robustness. In this scenario, attackers need to consider semantic or phonetic similarity to guarantee unnoticeable perturbation. Therefore, in addition to the commonly used optimization method used to attack a seq2seq translation model [86], some heuristic approaches are proposed to find substitute words that attack RNN-based dialogue generation models [268].

3.1.3 Defense Strategies. Under evasion adversarial attacks, there are different types of countermeasures to prevent the adversary from creating harmful effects. During the training stage, *adversarial training* aims to train a robust model by using adversarial samples during the training process. *Certified defense* works to achieve robustness over all perturbations within a certain bound. For defenses that happen at inference time, *adversarial example detection* aims to distinguish adversarial examples, so users can reject the prediction of harmful examples. To defend against attacks happening during training, feature extractors trained on auxiliary data have been used to extract useful features and correspondingly avoid shortcuts that can help prevent poisoning attack [376]. *Adversarial training* has also been shown to be effective against poisoning attacks [181, 335] .

3.2 Representative Attack Methods

In this subsection, we introduce representative attack methods from two aspects: evasion attacks and poisoning attacks.

3.2.1 Evasion Attack. Evasion attacks occur at test time. When attackers want to generate an adversarial example for a test example, they need to find a distance measure to guarantee the perturbation size is small so that evasion attacks can be further categorized by how to constraint the perturbation, i.e., pixel constrained adversarial examples with a fixed l_p norm bound and adversarial examples under other types of constraints.

- **L_p bound attacks.** To guarantee the perceptual similarity between the adversarial example and the natural example, the perturbation is normally constrained within an l_p norm bound around the natural example. To find such perturbation, Projected Gradient Descent (PGD) adversarial attack [243] tries to calculate the adversarial example x' that maximizes the loss function:

$$\begin{aligned} &\text{maximize } \mathcal{L}(\theta, x') \\ &\text{subject to } ||x' - x||_p \leq \epsilon \text{ and } x' \in [0,1]^m \end{aligned}$$

  Here, ϵ is the perturbation budget and m is the dimension of the input sample x. This local maximum is calculated by the projected gradient ascent algorithm. At each time step, a small gradient step is made towards the direction of increasing loss value before the new example is projected back to the L_p norm bound. A representative attack method to achieve a standardized evaluation of robustness is called Autoattack [104], which is a strong evasion attack conducted by assembling four attacks, including three white-box attacks and one black-box attack. This brings a more reliable evaluation of model robustness under evasion attacks.
  There are also evasion attacks with special goals. The work [263] devises an algorithm that successfully misleads a classifier's decision on almost all test images. It tries to find a perturbation δ under a ϵ constraint satisfying that for any sample x from the test distribution, such a perturbation δ can misguide the classifier to give wrong decisions on most of the samples.
- **Beyond l_p bound attacks.** Recently people started to realize that l_p norm perturbation budget is neither sufficient to cover real-world noise nor a perfect measurement for perceptual similarity. Some studies seek to find the minimal perturbation necessary to change the class of a given input with respect to the l_p norm [103]. Other works propose different perturbation measurements, e.g., the Wasserstein distance [358, 360], to measure the changes of pixels.

3.2.2 Poisoning Attacks. As we introduced, poisoning attacks allow adversaries to take control of the training process.

- **Training Time Attack.** In the training time attack setting, perturbation only happens during the training time. For example, the "poisoning frog" attack inserts an adversarial image with the true label to the training set, to make the trained model wrongly classify target test samples [313]. It generates the adversarial example x' by solving the following problem:

$$x' = \text{argmin}_x \|Z(x) - Z(x_t)\|_2^2 + \beta \|x - x_b\|_2^2.$$

  Here, $Z(x)$ is the logits of the model for samples x, x_t and x_b are the samples from the target class and the original class, respectively. The result x' would be similar to the base class in the input space while sharing similar predictions with the target class. As a concrete example, the features of birds are intentionally added into

training samples of cats but the cats are still labeled as cats in training. As a consequence, it would mislead the model's prediction of other bird images that also contain bird features.

- **Backdoor Attack.** A backdoor attack requires that perturbations happen in both training and test data. A backdoor trigger only known by the attacker is inserted into the training data to mislead the classifier into giving a target prediction on all the test examples that contain the same trigger [84]. This type of attack is particularly dangerous because the model behaves normally on natural samples, which makes it even hard to notice the dark side of the model.

3.3 Representative Defense Methods

In this subsection, we introduce representative defense methods from the aforementioned categories.

3.3.1 Robust Optimization / Adversarial Training. Adversarial training aims to train models that give resistant predictions to adversarial examples. The training objective is formulated as a min-max problem that tries to minimize the error risk on the maximum adversarial loss within a small area around the training data samples [349]. With this bi-level optimization process, the model achieves partial robustness but still suffers from longer training time, natural accuracy and robust accuracy trade-offs, and robust overfitting issues. There are several studies making efforts to improve standard adversarial training from different perspectives. In [342], the trade-off issue is revealed. TRADES [386] takes a step toward balancing the natural accuracy and robust accuracy by adding a regularization term to minimize the prediction difference between adversarial samples and natural samples. Other works [314, 357] boost the training speed by estimating the gradient of the loss by treating the parameters of the last few model layers as constant layers when generating adversarial samples. They can shorten the training time to one-fourth of the GPU time – or even shorter with comparable robust performance. To mitigate robust overfitting, different classic techniques, such as early stop, weight decay, and data augmentations have been investigated [290, 359]. It is evident from recent work [71] that using data augmentation methods is a promising direction to further boost adversarial training performance.

3.3.2 Certified Defense. Certified defense seeks to learn provably robust DNNs against specific norm-bounded perturbations [286, 356]. In empirical defenses, such as adversarial training, robustness is achieved to a certain extent; in certified robust verification, however, we want to exactly answer the question of whether we can find an adversarial example for a given example. For instance, a randomized smoothing based classifier [95] aims to build an absolutely smooth classifier by making decisions according to the majority of predictions of all neighborhood examples around the original test example. Achieving such smoothness requires considerably greater computation resources, which is a challenge in practice.

3.3.3 Detection. In order to distinguish the adversarial examples in data distribution and to prevent the harmful effect, people design detection algorithms. A common way to achieve detection is to build another classifier to predict whether a sample is adversarial or not. The work [158] trains a binary classification model to discriminate all adversarial examples apart from natural samples and then builds ML models on recognized natural samples. Other works detect the adversarial samples based on the statistic property of adversarial sample distribution difference compared to natural sample distribution. Other work [162] uses a statistical test, i.e., a Maximum Mean Discrepancy (MMD) test, to determine whether two datasets are drawn from the same distribution. It uses this tool to test whether a group of data points are natural or adversarial; however, it is shown in [68] that evasion adversarial examples are not easily detected. This paper covers ten detection methods and finds out that those defenses lack thorough security evaluations and there is no clear

evidence to support that adversarial samples are intrinsically different from clean samples. Recent work [82] proposes a new direction for black-box adversarial detection. It detects the attacker's purpose based on the historical queries and sets a threshold for the distance between two input image queries to detect suspicious attempts to generate adversarial examples.

3.4 Applications in Real Systems

When machine learning is applied to real-world, safety-critical tasks, the existence of adversarial examples becomes more dangerous and may cause severe consequences. In the following section, we illustrate the potential threats from adversarial examples to real-world applications in different domains.

3.4.1 Image Domain. In the autonomous driving domain, road sign detection is an important task. However, with some small modifications, the road sign detection system [130, 324] in the vehicle would recognize 35 mph as 85 mph and cannot successfully detect a stop sign as shown in Figure 3. Deep learning is also widely applied in authentication tasks. An attacker can wear a special glass to pretend to be an authorized identity in order to mislead the face recognition model; this deceit can be accomplished by labeling a few face samples that wear certain glasses as the target identity and by inserting those mislabeled samples into the training set [84]. Another example of identity threat is that the person detection system for images or videos can also be broken by wearing an adversarial T-shirt [364].

Fig. 3. The stop sign could not be distinguished by machines with modifications.

3.4.2 Text Domain. Adversarial attacks also happen in natural language processing tasks, such as text classification, machine translation, and dialogue generation. For machine translation, sentence and word paraphrasing on input texts are conducted to craft adversarial examples [220]. The attacker first builds a paraphrasing corpus that contains a lot of words and sentence paraphrases. To find an optimal paraphrase of an input text, a greedy method is adopted to search for valid paraphrases for each word or sentence from the corpus. Moreover, it proposes a gradient-guided method to improve the efficiency of the greedy search. Another work [230] treats the neural dialogue model as a black box and adopts a reinforcement learning framework to effectively find trigger inputs for targeted responses. This type of black-box setting is stricter but more realistic, while the requirements for the generated responses are properly relaxed. The generated responses are expected to be semantically identical to the targeted ones but not necessarily exactly match them.

3.4.3 Audio Data. The state-of-art speech-to-text transcription networks, such as DeepSpeech [168], can be attacked by a small perturbation [70]. Given any speech waveform x, an inaudible sound perturbation δ is added to make the synthesized speech $x + \delta$ recognized as any targeted desired phrase. In this work [303], authors propose an adversarial attack method against the YouTube CopyRight detection system to avoid music with copyright issues being detected during uploading. The attack uses a neural network to extract features from the music piece to create a fingerprint, which is used for checking whether the music matches existing copyrighted music. A gradient-based adversarial attack on the original audio piece can create a large difference in the output fingerprint such that the modified audio can successfully avoid the detection of YouTube's CopyRight detection system.

3.4.4 Graph Data. Zügner et al. [404] considers attacking node classification models, graph convolutional networks [208], by modifying the node connections or node features. In this setting, an adversary is allowed to add or remove edges between nodes or change the node features with a limited number of operations in order to mislead the GCN model that is trained on the perturbed graph. The work [405] attempts to poison the graph so that the global node classification performance of GCN will drop – and even be made almost useless. They optimize the graph structure as the hyper-parameters of the GCN model with the meta-learning technique. The goal of this attack [44] is to perturb the graph structure to corrupt the quality of node embedding, affecting the downstream tasks' performance, including node classification or link prediction. Pro-GNN [191] tries to jointly learn a structural graph and a graph neural network from the perturbed graph guided by some intrinsic properties of a real-world graph, such as low-rank and feature smoothness. The defense method could learn a clean adjacency matrix close to the perturbed one and limit its norm to guarantee the low-rank property. Meanwhile, a feature smoothness regularizer is also utilized to penalize rapid changes in features between adjacent nodes. Then a robust GNN would be built upon the learned graph.

3.5 Robustifying Real World System

To make real-world ML systems safer, many strategies can be applied. As we mentioned, the Youtube music copyright system can be fooled by simply adding some unnoticeable noise. With the help of adversarial training, the problem could be eased [303]. Also, A chatbot is a commonly used machine learning system, and inappropriate language can be a safety problem. To mitigate the harmful effect, the work [366] proposes a robust evaluation and enhancement method to improve the reliability of dialogue systems. Meanwhile, people are seeking methods to create adversarial autonomous driving environments to enhance the safety of autonomous driving.

3.6 Surveys and Tools

In this subsection, we list related resources about adversarial robustness, including surveys and tools.

3.6.1 Surveys. Xu et al. [364] gives a comprehensive introduction of concepts and covers representative attack and defense algorithms in different domains, including image classification, graph classification, and natural language processing. For the surveys in a specific domain, Akhtar and Mian [25] provides a comprehensive introduction to adversarial threats in a computer vision domain [76]; Jin et al. [190] gives a thorough review of the latest adversarial robustness techniques in the graph domain; and Zhang et al. [389] focuses on natural language processing and summarizes important algorithms on adversarial robustness in the text domain.

3.6.2 Tools. *Advertorch* [118] is a Pytorch toolbox containing popular attack methods in the image domain. *DeepRobust* [226] is a comprehensive and up-to-date adversarial attacks and defenses library based on Pytorch that includes not

only algorithms in the image domain but also the graph domain. This platform provides convenient access to different algorithms and evaluation functions to illustrate the robustness of image classification models or graph properties. *RobustBench* [102] provides a robust evaluation platform by the Autoattack algorithm for different adversarial training models. This platform also provides well-trained robust models by different adversarial training methods, which can save resources for researchers. A summary of the above-mentioned open-source toolkits and frameworks on robustness in AI can be found in Table 2.

Table 2. Representative open-source toolkits and frameworks on robustness in AI.

Toolkit	Characteristics
Cleverhans [275]	A Tensorflow library for benchmarking the vulnerability of machine learning models to adversarial examples.
Advertorch [118]	A Pytorch toolbox containing attack methods for image classification.
DeepRobust [226]	A Pytorch toolbox containing attack and defense methods in image and graph domain.
RobustBench [102]	A tool to standardize the evaluation of adversarial robustness.

3.7 Future Directions

For adversarial attacks, people are seeking more general attacking methods to evaluate adversarial robustness. For black-box attacks, efficiently generating adversarial examples with fewer adversarial queries is often challenging. For training in adversarial defense methods, an important issue is the robust overfitting and lack of generalization in both adversarial and natural examples. These problems remain unsolved and need further improvements. Another direction for adversarial training is to build robust models against more general adversarial examples, including but not limited to different l-p bound attacks. For certified defenses, one possible direction is to train a model with robust guarantees more efficiently, since the current certified defense methods require a large number of computational resources.

4 NON-DISCRIMINATION & FAIRNESS

As AI plays an increasingly irreplaceable role in various scenarios closely related to people's vital interests, such as recidivism prediction, financial risk assessment and job recommendation, an AI system ought to avoid discriminatory behaviors in human-machine interactions and ensure fairness in decision making for any individuals or groups. Otherwise, it would lose the trust from various stakeholders. With the rapid spread of AI in our daily lives, more and more evidence indicates that AI systems show human-like discriminatory bias or make unfair decisions. For example, a recidivism prediction software used by U.S. courts often assigns a higher risky score for an African American than a Caucasian with a similar profile[1] [253]; a job recommendation system promotes more STEM employment opportunities to male candidates than to females [215]. Moreover, Tay, the online AI chatbot developed by Microsoft, produced a lot of improper racist and sexist comments, which led to its closure within 24 hours of release [354]; dialogue models trained on human conversations show bias toward females and African Americans by generating more offensive and negative responses for these groups [229]. Fairness in AI demands considerable attention. Recently, many works have emerged to define, recognize, measure, and mitigate the bias in AI algorithms. In this section, we aim to give a comprehensive overview of the cutting-edge research progress addressing fairness issues in AI. In the subsections, we first present concepts and definitions regarding fairness in AI. We then provide a detailed taxonomy to discuss

[1]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

different origins of algorithmic bias, different types of bias, and fairness. We then review and classify popular bias mitigation technologies for building fair AI systems. Next, we introduce the specific bias issues and the applications of bias mitigation methods in real-world AI systems. In this part, we categorize the works according to the types of data processed by the system. Finally, we discuss the current challenges and future opportunities in this field. We expect that researchers and practitioners can gain a sense of direction and understanding from a broad overview of bias and fairness issues in AI and deep insight into the existing solutions, so as to advance progress in this field.

4.1 Concepts and Taxonomy

Before we go deep into nondiscrimination and fairness in AI, we need to understand how relative concepts, such as bias and fairness, are defined in this context. In this subsection, we briefly illustrate the concepts of bias and fairness, and provide a taxonomy to introduce different sources of bias, different types of bias, and fairness.

4.1.1 Bias. In the machine learning field, the word "bias" has been abused. It conveys different meanings in different contexts. We first distinguish the concept of bias in the context of AI non-discrimination and fairness from that in other contexts. There are three categories of bias: **productive bias**, **erroneous bias**, and **discriminatory bias**. Productive bias exists in all machine learning algorithms. It is beneficial and necessary for an algorithm to be able to model the data and make decisions [175]. Based on the "no free lunch theory" [355], only a predictive model biased toward certain distributions or functions can achieve better performance on modeling them. Productive bias helps an algorithm to solve certain types of problems. It is introduced through our assumptions about the problem, which is specifically reflected as the choice of a loss function, an assumed distribution, or an optimization method, etc. Erroneous bias can be viewed as a systematic error caused by faulty assumptions. For example, we typically assume that the distribution of the training data is consistent with the real data distribution. However, due to selection bias [244] or sampling bias [253], the collected training data may not be able to represent the real data distribution. Thus, the violation of our assumption can lead to the learned model's undesirable performance on the test data. Discriminatory bias is the kind of bias we are interested in under AI nondiscrimination and fairness. As opposed to fairness, discriminatory bias reflects an algorithm's unfair behaviors toward a certain group or individual, such as producing discriminatory content or performing less well for some people [316]. In the rest of this paper, when we mention bias, we refer to discriminatory bias.

Sources of Bias. The bias in an AI system can be produced by different sources, namely, the data, the algorithm, or the evaluation method. Bias within data comes from different phases of data generation, from data annotation to data collection and data processing [272, 315]. In the phase of data annotation, bias can be introduced due to a non-representative group of annotators [192], inexperienced annotators [278], or preconceived stereotypes held by the annotators [309]. In the phase of data collection, bias can emerge due to the selection of data sources or how data from several different sources are acquired and prepared [272]. In the data processing stage, bias can be generated due to data cleaning [115], data enrichment [96], and data aggregation [343].

Types of Bias. Bias can be categorized into different classes from different perspectives. It can be explicit or implicit. *Explicit bias*, also known as direct bias, occurs when the sensitive attribute explicitly causes an undesirable outcome for an individual; while *implicit bias*, also known as indirect bias, indicates the phenomenon that an undesirable outcome is caused by nonsensitive and seemingly neutral attributes, which in fact have some potential associations with the sensitive attributes [387]. For example, the residential address seems a nonsensitive attribute, but it can correlate with the race of a person according to the population distribution of different ethnic groups [387]. Moreover, language

style can reflect the demographic features of a person, such as race and age [182, 231]. Bias can be acceptable and unacceptable. *Acceptable bias*, also known as explainable bias, describes a situation where the discrepancy of outcomes for different individuals or groups can be reasonably explained by factors. For example, models trained on the UCI Adult dataset predict higher salaries for males than females. Actually, this is because males work for a longer time per week than females [200]. Based on this fact, such biased outcomes are acceptable and reasonable. Conversely, bias that cannot be explained appropriately is treated as *unacceptable bias*, which should be avoided in practice.

4.1.2 Fairness. The fairness of an algorithm is defined as "the absence of any prejudice or favoritism toward an individual or a group based on their intrinsic or acquired traits in the context of decision making" [253, 310]. Furthermore, according to the object of the study, fairness can be further defined as group fairness and individual fairness.

Group Fairness. Group fairness requires that two groups of people with different sensitive attributes receive comparable treatments and outcomes statistically. Based on this principle, various definitions have been proposed, such as Equal Opportunity [169], which requires people from two groups to be equally likely to get a positive outcome when they indeed belong to the positive class; Equal Odds [169], which requires that the probability of being classified correctly should be the same for different groups; and Demographic Parity [123], which requires different groups to have the same chance to get a positive outcome, etc.

Individual Fairness. While group fairness can maintain fair outcomes for a group of people, a model can still behave discriminatorily at the individual level [123]. Individual fairness is based on the understanding that similar individuals should be treated similarly. A model satisfies individual fairness if it gives similar predictions to similar individuals [123, 213]. Formally, if individuals i and j are similar under a certain metric δ, the difference between the predictions given by an algorithm M on them should be small enough: $|f_M(i) - f_M(j)| < \epsilon$, where $f_M(\cdot)$ is the predictive function of algorithm M that maps an individual to an outcome, and ϵ is a small constant.

4.2 Methods

In this subsection, we introduce bias mitigation techniques. Based on which stage of an AI pipeline is to interfere, the debiasing methods can be categorized into three types: **pre-processing**, **in-processing** and **post-processing** methods. Representative bias mitigation methods are summarized in Table 3.

Table 3. Representative debiasing strategies in the three categories.

Category	Strategy	References
Pre-processing	Sampling	[20, 35, 394]
	Reweighting	[64, 199, 385]
	Blinding	[79, 94, 170, 380]
	Relabelling	[101, 164, 199]
	Adversarial Learning	[19, 141, 195]
In-processing	Reweighting	[188, 211]
	Regularization	[22, 139]
	Bandits	[126, 236]
	Adversarial Learning	[75, 231, 232, 384]
Post-processing	Thresholding	[169, 184, 254]
	Transformation	[89, 205, 266]
	Calibration	[172, 206]

Pre-processing Methods. Pre-processing approaches try to remove the bias in the training data to ensure the fairness of an algorithm from the origin [199]. This category of methods can be adopted only when we have access to the training data. Various strategies are proposed to interfere with training data. Specifically, Celis et al. [74] propose to adaptively sample the instances that are both diverse in features and fair to sensitive training attributes. Moreover, reweighting methods [198, 385] try to mitigate the bias in training data by adaptively up-weighting the training instances of underrepresented groups, while down-weighting those of overrepresented groups. Blinding methods try to make a classifier insensitive to a protected variable. For example, Hardt et al. [170] force a classifier to have the same threshold value for different race groups to ensure that the predicted loan rate is equal for all races. Some works [198, 382] try to relabel the training data to ensure the proportion of positive instances are equal across all protected groups. Additionally, Xu et al. [362] take advantage of a generative adversarial network to produce bias-free and high-utility training data.

In-processing Methods. In-processing approaches address the bias at the algorithm level and try to eliminate bias during the model training process. They often seek to create a balance between performance and fairness [73]. Krasanakis et al. [211] propose an in-processing re-weighting approach. They first train a vanilla classifier to learn the weights of samples and then retrain the classifier using these weights. Some works [154, 201] take advantage of regularization methods, where one or more penalty terms are added into the objective function to penalize biased outcomes. The idea of adversarial learning is also adopted in in-processing debiasing methods. Liu et al. [232] design an adversarial learning framework to train neural dialogue models that are free from gender bias. Alternatively, bandits recently have emerged as a novel idea for solving fairness problems. For example, Joseph et al. [193] propose solving the fairness problem under a stochastic multi-armed bandit framework with fairness metrics as the rewards and the individuals or groups under investigation as the arms.

Post-processing Methods. Post-processing approaches directly make transformations on the model's outputs to ensure fair final outcomes. Hardt et al. [170] propose approaches to determine threshold values via measures such as equalized odds specifically for different protected groups to find a balance between the true and false positive rates to minimize the expected classifier loss. Feldman et al. [138] propose a transformation method to learn a new fair representation of the data. Specifically, they transform the SAT score into a distribution of the rank order of the students independent of gender. Pleiss et al. [279] borrow the idea of calibration to build fair classifiers. Similar to the traditional definition of calibration that the proportion of positive predictions should be equal to the proportion of positive examples, they force the conditions to hold for different groups of people. Nevertheless, they also find that there is a tension between prediction accuracy and calibration.

4.3 Applications in Real Systems

In this subsection, we summarize the studies regarding bias and fairness issues in real-world AI systems during different tasks. We introduce the works following the order of different data domains, including tabular data, images, texts, audios, and graphs. For each domain, we describe several representative tasks and present how AI systems can be biased on these tasks. A summary of the representative works can be found in Table 4. In addition, we demonstrate various examples of adopting the debiasing methods from Section 4.2 to mitigate the bias in real-world AI systems.

4.3.1 Tabular Domain. Tabular data is the most common format of data in machine learning; thus, the research on bias in machine learning is predominantly conducted on tabular data. In the recent decade, researchers have investigated how algorithms can be biased in classification, regression, and clustering tasks. For classification, researchers find evidence

Table 4. A summary of bias detection works in different data domains.

Domain	Task	References
Tabular Data	Classification	[62, 63, 155, 170, 197, 255]
	Regression	[21, 38]
	Clustering	[31, 83]
Image Data	Image Classification	[273]
	Face Recognition	[59, 178]
	Object Detection	[302]
Text Data	Text Classification	[50, 119, 182, 209, 277, 385]
	Embedding	[46, 56, 156, 246, 274, 396]
	Language Modeling	[48, 148, 238, 320, 371]
	Machine Translation	[34, 91, 157, 329, 345]
	Dialogue Generation	[106, 117, 229]
Audio Data	Speech Recognition	[72, 178, 294, 336]
Graph Data	Node Embedding	[51]
	Graph Modeling	[107]

that machine learning models for credit prediction [197] and recidivism prediction [93] tasks can show significant prejudice toward certain demographic attributes of a person, such as race and gender. Berk et al. [38] and Agarwal et al. [21] investigate multiple regression tasks, from salary estimation to crime rate prediction, showing unfair treatment for different races and genders. Backurs et al. [31] and Chen et al. [83] discover unfairness in clustering algorithms with a belief that as data points, different groups of people are entitled to be clustered with the same accuracy.

4.3.2 Image Domain. Machine learning models in computer vision have also shown unfair behaviors. In [59, 178], the authors show that face recognition systems work better for white compared to darker faces, and also show similar bias in terms of gender. An image classification application developed by Google has been accused of labeling black people as "gorillas" [273]. In [302], the authors discover the difference in the performances of smiling detection models on different genders and races. The work [397] tackle the social bias in visual semantic role labeling, e.g., associating cooking roles with women. They introduce corpus-level constraints for calibrating existing structured prediction models. In the work [349], a visual recognition benchmark is designed for studying bias mitigation.

4.3.3 Text Domain. A large number of works have shown that algorithmic bias exists in various natural language processing tasks. Word embeddings often exhibit a stereotypical bias for text data, causing a serious risk of perpetuating problematic biases in imperative societal contexts. In [47], the authors first shows that popular state-of-the-art word embeddings regularly mapped men to working roles and women to traditional gender roles, leading to significant gender bias which is even inherited in downstream tasks. Following the research of word embeddings, the same patterns of gender bias are discovered in sentence embeddings [247]. In the task of co-reference resolution, researchers demonstrate in [398] that rule-based, feature-based, and neural network-based co-reference systems all show gender bias by linking gendered pronouns to pro-stereotypical entities with higher accuracy than anti-stereotypical entities. Language models can also learn gender discrimination from man-made text data [49], which tend to generate certain words reflecting gender stereotypes with different probabilities in the context of males and females. As for machine translation, it has been illustrated that Google's translation system suffers from gender bias by showing favoritism toward males for stereotypical fields, such as STEM jobs when translating sentences taken from the U.S. Bureau of Labor Statistics into a

dozen gender-neutral languages [281]. Dialogue systems, including generative models and retrieval-based models, also show bias toward different genders and races by producing discriminatory responses [229, 232].

4.3.4 Audio Domain. Voice recognition systems show gender bias by processing the voices of men and women differently [178]. It is found that medical voice-dictation systems recognize voice inputs from males versus females with higher accuracy [294]. It is shown in [72] that voice control systems on vehicles worked better for males than females. Google's speech recognition software can understand queries from male voices more consistently than those from females [336].

4.3.5 Graph Domain. AI applications on graph-structured data are ubiquitous in the real world. The fairness issues in these problems are drawing increasing attention from researchers. Existing graph embedding techniques can learn node representations correlated with protected attributes, such as age and gender. Consequently, they exhibit bias toward certain groups in real-world applications, like social network analysis and recommendations [51]. Graph neural networks (GNNs) also inherit bias from training data and even magnify the bias through GNN's graph structures and message-passing mechanism [107].

4.3.6 Debiasing Real Systems. Many attempts have been made to ensure fairness in real-world AI systems. Zhang et al. [384] utilizes the in-processing adversarial debiasing strategy to train a fair classification model. During the training process, an adversary is introduced to judge whether the outcomes are fair or not, and the feedback from the adversary will serve as the guidance to improve the fairness of the classifier. The work [178] introduces a pre-processing method where training data is balanced to alleviate gender bias in face recognition tasks. In terms of text data, Zhang et al. [385] developes an instance weighting framework that can effectively mitigate the bias of text classifiers toward demographic identity-terms, while the prediction capability of the models is not significantly affected.

4.4 Surveys and Tools

In this subsection, we gather the existing surveys, tools and repositories on fairness in AI to facilitate readers wishing to explore this field further.

4.4.1 Surveys. The problem of fairness has been studied in multiple disciplines other than computer science for more than a half century. In one survey [183], the authors trace the evolution of the notions and measurements of fairness in different fields, such as education and hiring, over the past 50 years. They provide a comprehensive comparison between the past and current definitions to encourage a deeper understanding of modern fairness in AI. Zliobaite [402] provides an early survey on measuring indirect discrimination in machine learning. In this survey, the authors review early approaches for measuring bias in data and predictive models. They also analyze the measurements from other fields and explore the possibility of their use in machine learning. Corbett-Davies and Goel [99] provide a critical review on the measurements of fairness, showing the limitations of the existing fairness criteria in classification tasks in machine learning. Mehrabi et al. [253] contribute a comprehensive survey on bias and fairness in machine learning. In this survey, the authors provide a detailed taxonomy of the bias and fairness definitions in machine learning, and also introduce the bias observed in the data and algorithms in different domains of AI and the state-of-the-art debiasing methods. Caton and Haas [73] provide an overview of the existing debiasing approaches for building fair machine learning models. They organize the extant works into three categories and 11 method areas and introduce them following their taxonomy. Moreover, there are some surveys regarding bias and fairness in specific domains of AI. Blodgett et al. [43] review the papers analyzing bias in NLP systems, providing critical comments on such works and indicating that many

existing works suffer from unclear and inconsistent motivations and irrational reasoning. They also offer suggestions to normalize future studies on bias in NLP. Chen et al. [80] summarize and organize the works on bias and debias in recommender systems, and discuss future directions in this field.

4.4.2 Tools. In recent years, some organizations or individual researchers have provided multi-featured toolkits and repositories to facilitate fair AI. The repository *Responsibly* [237] collects the datasets and measurements for evaluating bias and fairness in classification and NLP tasks. The project *FairTest* [339] provides an unwarranted associations (UA) framework to discover unfair user treatment in data-driven algorithms. *AIF360* [36] collects popular datasets for fairness studies and provides the implementations of common debiasing methods for binary classification. *Aequitas* [307] is released as an audit toolkit to test the bias and fairness of models for multiple demographic groups on different metrics. The repository *Fairness Measurements* provides datasets and codes for quantitatively measuring discrimination in classification and ranking tasks. A summary of the above-mentioned open-source toolkits and frameworks on fairness in AI can be found in Table 5.

Table 5. Representative open-source toolkits and frameworks on fairness in AI.

Toolkit	Characteristics
Responsibly [237]	Datasets and measurements for bias and fairness evaluation on classification tasks and NLP tasks.
FairTest [339]	A framework for discovering unfair user treatment in data-driven algorithms.
AIF360 [36]	Popular datasets for fairness research and implementations of common debiasing algorithms.
Aequitas [307]	An audit framework for testing the fairness of ML models.
Fairness Measurements [2]	Datasets and implementations for measuring bias in classification and ranking tasks.

4.5 Future Directions

Fairness research still possesses a number of outstanding challenges.

- **Trade-off between fairness and performance.** Studies on fairness in different fields have confirmed the existence of the trade-off between fairness and performance of an algorithm [39, 100, 283]. The improvement of the fairness of an algorithm typically comes at the cost of performance degradation. Since both fairness and performance are indispensable, extensive research is needed to help people better understand an algorithm's trade-off mechanism between them, so that practitioners can adjust the balance in practical usage based on the actual demand;
- **Precise conceptualization of fairness.** Although extensive research has been conducted on bias and fairness in AI, too much of this work formulates its concerns under a vague concept of bias that refers to any system harmful to human behaviors but fails to provide a precise definition of bias or fairness specific to their setting [43]. In fact, different forms of bias can appear in different tasks, even in the same task. For example, in a recommender system, popularity bias can exist toward both the users and items [80]. In a toxicity detection algorithm, race bias can exist toward both the people mentioned in texts and the authors of texts [231]. To study any fairness problem, a precise definition of bias indicating how, to whom, and why an algorithm can be harmful must be articulated. In this way, we can make the research on AI fairness in the whole community more standardized and systematic;

- **From equality to equity.** Fairness definitions are often associated with equality to ensure that an individual or a conserved group, based on race or gender, are given similar amounts of resources, consideration, and results. Nonetheless, the area of equity has been heavily under-examined [253], where this notion pertains to the particular resources for an individual or a conserved group to be successful [159]. Equity remains an interesting future direction, since the exploration of this definition can extend or contradict existing definitions of fairness in machine learning.

5 EXPLAINABILITY

Making the AI techniques transparent can enhance the trustworthiness, where humans can understand how/why it can work well and when/why/where it doesn't work, so that users can adopt the AI techniques accordingly and fully trust them afterward. The improved predictive performance of AI systems has often been achieved through increased model complexity [120, 262]. A prime example is the paradigm of deep learning, dominating the heart of most state-of-the-art AI systems. However, deep learning models are treated as black-boxes, since most of them are too complicated and opaque to be understood and are developed without explainability [227]. More importantly, without explaining the underlying mechanisms behind the predictions, deep models cannot be fully trusted, which prevents their use in critical applications pertaining to ethics, justice, and safety, such as healthcare [258], autonomous cars [221], and so on. Therefore, building a trustworthy AI system requires an understanding of how particular decisions are made [144], which has led to the revival of the field of eXplainable Artificial Intelligence (XAI). In this section, we aim to provide an intuitive understanding and high-level insights into the recent progress of explainable AI. First, we provide the concepts and taxonomy regarding explainability in AI. Second, we review representative explainable techniques for AI systems according to the aforementioned taxonomy. Third, we introduce real-world applications of explainable AI techniques. Finally, we provide some surveys and tools and discuss future opportunities for explainable AI.

5.1 Concepts and Taxonomy

In this subsection, we introduce the concepts of explainability in AI. We then provide a taxonomy of different explanation techniques.

5.1.1 Concepts. In the context of machine learning and AI literature, explainability and interpretability are usually used by researchers interchangeably [262]. One of the most popular definitions of explainability is the one from Doshi-Velez and Kim, who define it as "the ability to explain or to present in understandable terms to a human" [120]. Another popular definition is from Miller, who defines explainability as "the degree to which a human can understand the cause of a decision" [257]. In general, the higher the explainability of an AI system is, the easier it is for someone to comprehend how certain decisions or predictions have been made. Meanwhile, a model is better explainable than other models if its decisions are easier for a human to comprehend than those of others.

While explainable AI and interpretable AI are very closely related, subtle differences between them are discussed in some studies [153, 300, 391].

- A model is interpretable if the model itself is capable of being understood by humans on its predictions. When looking at the model parameters or a model summary, humans can understand exactly the procedure on how it made a certain prediction/decision and, even given a change in input data or algorithmic parameters, it is the extent to which humans can predict what is going to happen. In other words, such models are intrinsically

transparent and interpretable, rather than black-box/opaque models. Examples of interpretable models include decision trees and linear regression.

- An explainable model indicates that additional (post hoc) explanation techniques are adopted to help humans understand why it made a certain prediction/decision that it did, although the model is still black-box and opaque. Note that such explanations are often not reliable and can be misleading. Examples of such models would be deep neural network based models, where the models are usually too complicated for any human to comprehend.

5.1.2 Taxonomy. Techniques for AI's explanation can be grouped according to various criteria such as model, scope and method.

- **By Model: Intrinsic or Post-hoc.** If AI models are considered transparent or interpretable due to their simple structures, then these interpretable techniques are called a model intrinsic explanation. In contrast, post-hoc methods refer to the explainability methods that are developed to explain the target AI model after model training.
- **By Scope: Local or Global.** If the method provides an explanation only for a specific instance, then it is a local explanation; if the method explains the whole model, then it is a global explanation.
- **By Method: Gradient-based, Perturbation-based, or Others.** Most existing explainability methods can be categorized into two main classes: gradient-based and perturbation-based methods. If the techniques employ the partial derivatives on input instances to generate attributions, then these techniques are called a gradient-based explanation. If the techniques focus on the changes or modifications of input data, we name them a perturbation-based explanation. In addition, there are other explanation techniques beyond gradient/perturbation-based methods. For example, a counterfactual explanation usually refers to a causal situation in the form, "If X had not occurred, Y would not have occurred." In general, counterfactual explanation methods are post-hoc and can be used to explain predictions of individual instances (local) [262, 346].

5.2 Methods

In this subsection, we introduce some representative explanation techniques according to the aforementioned taxonomy. A summary of the representative works can be found in Table 6.

5.2.1 By Model: Model-intrinsic or Post-hoc. Any explainable algorithm that is dependent on the model architecture can fall into the model-intrinsic category. In contrast, post-hoc methods apply to any model for being generally applicable. In general, there are significant research interests in developing post-hoc methods to explain the predictions of an existing well-performing neural networks model. This criterion also can be used to distinguish whether interpretability is achieved by restricting the complexity of the AI model. Intrinsic interpretability refers to AI models that are considered interpretable (white-box) due to their simple model architecture, while most post-hoc explanations are widely applied into (black-box) deep neural networks which are highly complicated and opaque due to their millions of parameters.

- **Model-intrinsic Explanations.** The model in this category is often called an intrinsic, transparent, or white-box explanation. Generally, without designing an additional explanation algorithm, this type of interpretable technique cannot be re-used by other classifier architectures. Therefore, the model intrinsic methods of explanations are inherently model specific. Such commonly used interpretable models include linear/logistic regression, decision trees, rule-based models, Generalized Additive Models (GAMs), Bayesian networks, etc.

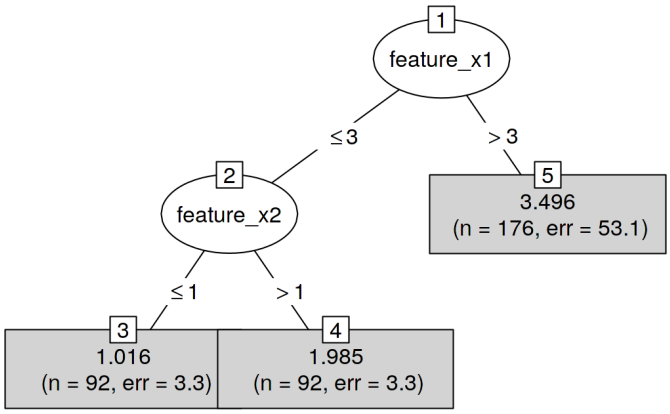
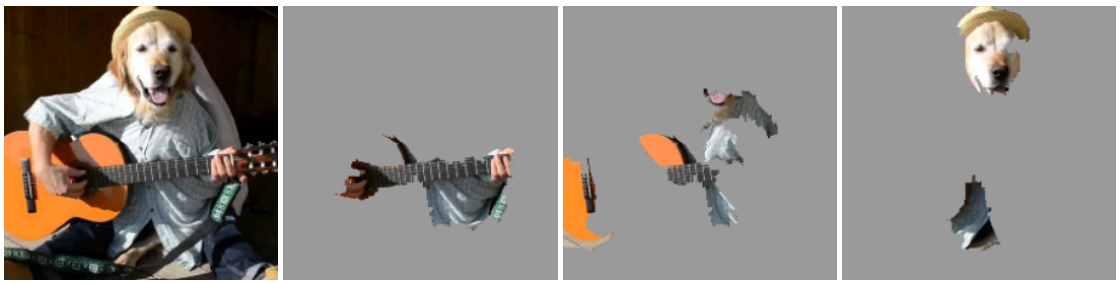


Fig. 4. The interpretation of decision tree is simple, where intermediate nodes in the tree represent decisions and leaf nodes can be class labels. Starting from the root node to leaf nodes can create good explanations on how the certain label is made by decision tree model. (Image Credit: [262])



(a) Original Image (b) Electric Guitar: 0.32 (c) Acoustic Guitar: 0.24 (d) Labrador: 0.21

Fig. 5. LIME explains deep neural networks on image classification task: top 3 predicted categories and the corresponding scores. (Image Credit: [289])

For example, the linear regression model [42], which is one of the most representative linear models in ML, aims to predict the target as a weighted sum of the feature of instances. With this linearity of the learned relationship, the linear regression model makes the estimation procedure simple and significantly understandable on a modular level (i.e., the weights) for humans. Mathematically, given one instance with d dimension of features $\mathbf{x}$, the linear regression model can be used to model the dependence of a predicted target $\hat{y}$ as follows:

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b = w_1 x_1 + ... + w_d x_d + b \tag{1}$$

where $\mathbf{w}$ and b denote the learned feature weights and the bias term, respectively. The predicted target $\hat{y}$ of linear regression is a weighted sum of its d dimension features $\mathbf{x}$ for any instance, where the decision-making procedure is easy for a human to comprehend by inspecting the value of the learned feature weights $\mathbf{w}$.
Another representative method is decision tree [284], which contains a set of conditional statements arranged hierarchically. Making predictions in a decision tree is also the procedure of explaining the model by seeking the path from the root node to leaf nodes (label), as illustrated in Figure 4.

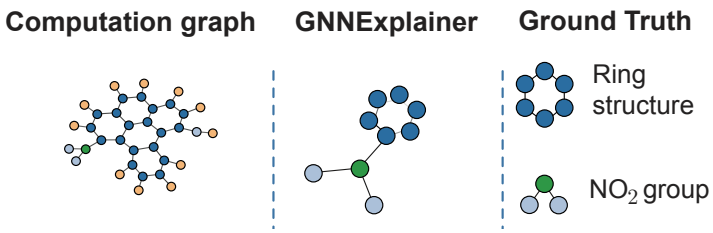


Fig. 6. GNNExplainer generates an explanation by identifying a small graph of the input graph for graph classification task on molecule graphs dataset (MUTAG). (Image Credit: [372])

Fig. 7. Image-specific class saliency maps were extracted using a single back-propagation pass through a DNN classification model. (Image Credit: [323])

- **Post-hoc Explanations.** The methods in this category are concerned with black-box well-trained AI models. More specifically, these methods do not try to create interpretable models, but to interpret already well-trained models. Such methods have been widely used for explaining complicated models, such as deep neural networks. That is why they sometimes are referred to as black-box explainability methods in the related scientific literature. One advantage of post-hoc methods over model-intrinsic ones is their flexibility. Post-hoc methods have also been applied in a variety of input modalities, such as images, text, graph-structured data, etc. Note that post-hoc methods can also be applied to intrinsically interpretable models.
  One of the most representative works in this category is Local Interpretable Model-Agnostic Explanations (**LIME**) [289]. For example, at the image domain, for any trained classifier, LIME is a proxy approach to randomly permute data by identifying the importance of local contiguous patches with similar pixels in a given instance and its corresponding label [289]. An illustrative example of LIME on a single instance for the top three predicted classes is shown in Figure 5.
  Additionally, to understand how any graph neural networks (GNNs) make a certain decision on graph-structured data, GNNExplainer learns soft masks for edges and node features to explain the predictions via maximizing the mutual information between the predictions of the original graph and those of the newly obtained graph [240, 372]. Figure 6 illustrates explanation examples generated by GNNExplainer for graph-structured data.

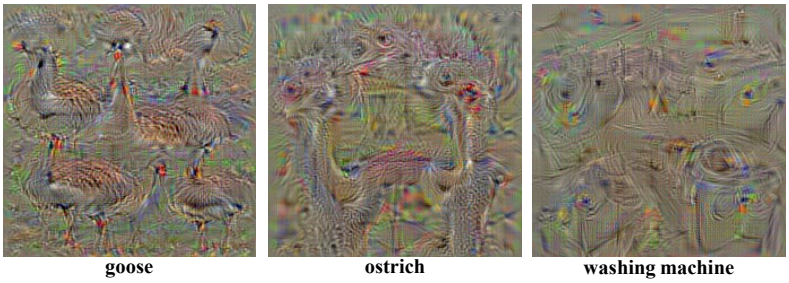


Fig. 8. Numerically computed images, illustrating the class appearance models. (Image Credit: [323])

5.2.2 By Scope: Local or Global. One important aspect of dividing the explainability techniques is based on the scope of explanation, i.e., local or global.

- **Local Explanations.** In general, the goal of locally explainable methods is to express the individual feature attributions of a single instance of input data x from the data population X. For example, given a text document and a model to understand the sentiment of text, a locally explainable model might generate attribution scores for individual words in the text.
  In the Saliency Map Visualization method [323], the authors compute the gradient of the output class category with regard to an input image. By visualizing the gradients, a fair summary of pixel importance can be achieved by studying the positive gradients that have more influence on the output [323]. An example of the class model is shown in Figure 7.
- **Global Explanations.** The goal of global explanations is to provide insights into the decision of the model as a whole and to have an understanding of attributions for a batch of input data or a certain label, not just for individual inputs. In general, globally explainable methods work on an array of inputs to summarize the overall behavior of the black-box model. Most linear, rule-based and tree-based models are inherently globally explainable. For example, conditional statements (intermediate nodes) in decision trees can give insight into how the model behaves in a global view, as shown in Figure 4.
  In terms of the DNNs models, Class Saliency Map Visualization [323] is trying to generate a particular image visualization by maximizing the score of class probability with respect to the input image (i.e., using the derivative of predicted class score with respect to the image). An example of the class model is shown in Figure 8. SP-LIME extends the vanilla LIME method (which provides local explanations on an individual instance) to give a global understanding via submodular optimization [289].

5.2.3 By Method: Gradient-based, Perturbation-based, or Others. This category is mainly defined by answering the question, "What is the algorithmic approach? Does it focus on the input data instance or the model parameters?" Based on the core algorithmic approach of the explanation method, we can categorize explanation methods as the ones that focus on the gradients of the target prediction with respect to input data, and those that focus on the changes or modifications of input data. Afterward, we also introduce other representative explanation techniques beyond gradient-based and perturbation-based methods.

- **Gradient-based Explanations.** In gradient-based methods, the explainable algorithm does one or more forward passes through the neural networks and generates attributions during the back-propagation stage utilizing partial

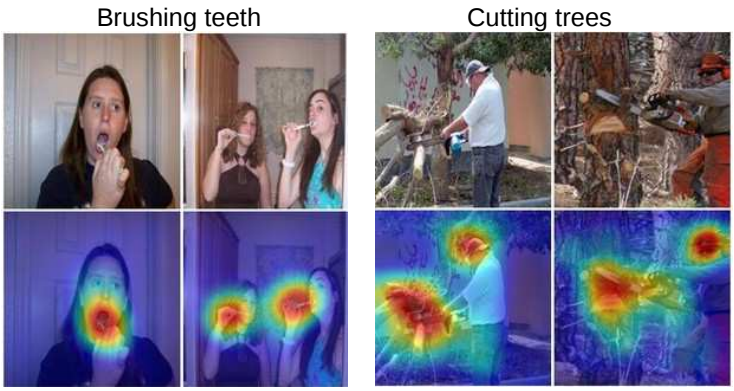


Fig. 9. Gradient-based Explanation: the CAM model produces class-specific regions of target images for visual explanation. (Image Credit: [399])

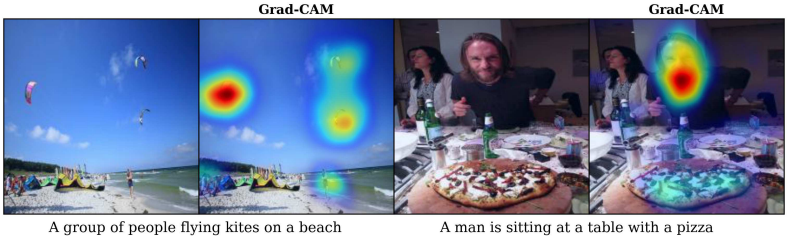


Fig. 10. Gradient-based Explanation: the Grad-CAM model localizes image regions considered to be important for producing the captions. (Image Credit: [312])

derivatives of the activations. This method is the most straightforward solution and has been widely used in computer vision to generate human-understandable visual explanations.

To understand how a CNN model makes decisions, Class activation mapping (CAM) [399] proposes modifying fully connected layers of the original CNN architecture using Global Average Pooling and generating important class-specific regions of the image for visual explanations via the forward passes process. An illustrative example is provided in Figure 9. Afterward, Gradient-weighted Class Activation Mapping (Grad-CAM) [312] generalizes the CAM model for any CNN model without requiring architectural changes or retraining and utilizes the gradient signal flowing into the final convolutional layer of a CNN for highlighting the important regions in the image. Figure 10 shows a visual explanation via Grad-CAM for an image captioning model.

- **Perturbation-based Explanations.** Perturbation-based explainable methods focus on variations in the input feature space to explain individual feature attributions toward the output class. More specifically, explanations are generated by iteratively probing a trained AI model with different variations of the inputs. These perturbations can be on a feature level by replacing certain features with zero or random counterfactual instances, picking one or a group of pixels (super-pixels) for the explanation, blurring, shifting, or masking operations, etc. In general, only a forward pass is sufficient to generate the attribution representations without the need for back-propagating

gradients. Shapley Additive explanations (SHAP) [239] visualizes feature interactions and feature importance by probing feature correlations by removing features in a game-theoretic framework.

- **Others.** Here, we present other explanation techniques that cannot be easily categorized into gradient/perturbation-based methods. For example, counterfactual explanations have been designed to answer hypothetical questions and describe how altering feature values of an instance would change the prediction to a predefined output [260, 262]. Taking the application for a credit card as an example, Peter gets rejected by AI banking systems and wonders why his application was rejected. To answer the question of "why," counterfactual explanations can be formulated as "What would have happened to this decision (from rejected to approved), if performing minimal changes in feature values (e.g., income, age, race, etc.)?" In fact, counterfactual explanations are usually human-friendly, since they are contrastive to the current instances and usually focus on a small number of features. To generate counterfactual explanations, Wachter et. al [346] propose Lagrangian style constrained optimization as follows:

$$\arg\min_{\hat{\mathbf{x}}} \max_{\lambda} \lambda \cdot (f(\hat{\mathbf{x}}) - \hat{y})^2 + d(\hat{\mathbf{x}}, \mathbf{x}) \tag{2}$$

where $\mathbf{x}$ is the original instance feature, and $\hat{\mathbf{x}}$ is the corresponding counterfactual input. $f(\hat{\mathbf{x}})$ is the predicted result of a classifier. The first term is the quadratic distance between the model prediction for the counterfactual $\hat{\mathbf{x}}$ and the targeted output $\hat{y}$. The second term indicates the distance $d(\cdot, \cdot)$ between the instance $\mathbf{x}$ to be explained and the counterfactual $\hat{\mathbf{x}}$, and λ is proposed to achieve the trade-off between the distance in prediction and the distance in feature values. Note that counterfactual explanations can be computed by gradients or perturbation methods.

Table 6. Summary of Published Research in Explainability of AI Systems.

Representative Models	Intrinsic vs. Post-hoc	Scope	Methodology
Linear model	Intrinsic	Global	Others
LIME [289]	Post-hoc	Local	Perturbation
SP-LIME [289]	Post-hoc	Global	Perturbation
CAM [399]	Post-hoc	Local	Others
Grad-CAM [312]	Post-hoc	Local	Gradient
SHAP [239]	Post-hoc	Both	Perturbation
Saliency Map Visualization [323]	Post-hoc	Local	Gradient
Class Saliency Map Visualization [323]	Post-hoc	Global	Gradient
GNNExplainer [372]	Post-hoc	Local	Perturbation
Surveys	[16, 29, 37, 110, 120, 121, 163, 189, 227, 257, 262, 338, 379, 391]		

5.3 Applications in Real Systems

When a prediction is complemented with explanations to understand how particular decisions are made, AI systems will gain more trust from humans. Various explanations techniques are proposed to integrate into real-world AI systems for enhancing the trustworthiness. In this subsection, we discuss representative real-world applications where explainability is crucial.

5.3.1 E-commerce. The explosive popularity of e-commerce sites has attracted an increasing number of people to spend more time shopping online. Recommender systems (RecSys) as intelligent systems have become increasingly

important in mitigating the information overload problem in e-commerce [134, 135]. These systems provide personalized information to help human decisions and have been widely used in various user-oriented online services [132], such as e-commerce item recommendations for everyday shopping (e.g., Amazon, Taobao), job recommendations for employment markets (e.g., LinkedIn), and friend recommendations to make people better connected (e.g., Facebook, Weibo) [131, 136]. Recently, in order to increase the trustworthiness, increasing attention has been paid to understanding why certain items have been recommended by deep learning-based recommender systems for end-users, because providing good explanations of personalized recommender systems can sufficiently motivate users to interact with items, help users make better and/or faster decisions, and increase users' trust in the intelligent recommender systems [241, 393]. For example, to achieve explainability in recommender systems, RuleRec [241] proposes a joint learning framework for accurate and explainable recommendations by integrating induction of several explainable rules from item association, such as, *Also view, Buy after view, Also buy, and Buy together*, where such explanations can sufficiently affect users online behaviors in e-commerce platforms.

5.3.2 Healthcare. Recently, AI techniques have demonstrated their significant impact on healthcare (e.g., drug discovery, disease diagnosis, gene analysis, etc.) in which decisions need to be carefully made based on solid evidence. For example, explainable AI has been proven to significantly accelerate the process of computer-assisted drug discovery [189, 344], such as molecular design, chemical synthesis planning, protein structure prediction, and macromolecular target identification. Explanations of graph neural networks have been conducted on a set of molecules graph-labeled for their mutagenic effect on the Gram-negative bacterium Salmonella typhimurium, with the goal of identifying several known mutagenic functional groups NH_2 and NO_2 [240, 372, 378]. A recent work [282] studies how the interpretation of filters within message-passing networks can lead to the identification of relevant toxicophore- and pharmacophore-like sub-structures for explainability, so as to help increase their reliability and foster their acceptance and usage in drug discovery and medicinal chemistry projects in the healthcare domain. In disease diagnosis, When AI models meet doctors who are experts in their sector for disease diagnosis, explanations should be made by considering the doctor's knowledge, application domain, the disease, patient conditions, etc. What's more, explanations should give doctors confidence in AI models' prediction that can lead to higher quality service [16].

5.3.3 Natural Language Processing. As one of the most broadly applied areas of AI, Natural Language Processing (NLP) investigates the use of computers to process or understand human (i.e., natural) languages [114]. Applications of NLP are everywhere, including dialogue systems, text summarization, machine translation, question answering, sentiment analysis, information retrieval, etc. Recently, deep learning approaches have obtained very promising performance across many different NLP tasks, which comes at the expense of models becoming less explainable [110, 265]. As a consequence, these NLP systems cannot be fully trusted, and humans might doubt how particular decisions are made. To address the issue, LIME [289] proposes to generate random input perturbations for a given document to explain the predicted categories for text classification in SVM models. CAML [265] employs an attention mechanism to select the segments that are most relevant for medical codes (ICD) from clinical text.

5.4 Surveys and Tools

In this subsection, we introduce existing surveys, tools and repositories on explainability in AI to facilitate the readers who wish to further explore this field.

5.4.1 Surveys. In the book [262], the author focuses on interpretable machine learning by introducing fundamental concepts to advanced interpretable models. For example, it first details related concepts of interpretability, followed by intrinsically interpretable models, such as linear regression, decision tree, rule-based methods, etc. Afterward, the book provides general post-hoc tools for interpreting black-box models and explaining individual predictions. Doshi-Velez et al. [120] raises the importance of intractability in machine learning and introduces a comprehensive survey at this field. There are surveys [29, 37, 121, 153, 163, 227] summarizing explanation approaches in machine learning. In addition, comprehensive surveys for specific applications also exist, such as recommender systems [391], medical information systems [338], natural language processing [110], graph neural networks [379], drug discovery [189], etc.

5.4.2 Tools. In this subsection, we introduce several popular toolkits that are open-sourced in the GitHub platform for explainable AI. *AIX360*[3] (AI Explainability 360) [30] is an open-source Python toolkit featuring state-of-the-art explainability methods and some evaluation metrics. Meanwhile, AIX360 also provides educational materials for non-technical stakeholders to quickly become familiar with interpretation and explanation methods. *InterpretML*[4] [270] is also an open-source python toolkit that exposes machine learning interpretability algorithms to practitioners and researchers. InterpretML exposes two types of interpretability: glass-box for machine learning models with model-intrinsic explanations, and black-box explainability techniques for explaining any existing AI systems. The package *DeepExplain* [27] mainly supports various gradient-based techniques and perturbation-based methods[5]. In addition, *DIG* [233] provides python toolkit for explaining graph deep learning[6]. A summary of the above-mentioned open-source toolkits and frameworks on explainable AI can be found in Table 7.

Table 7. Representative open-source toolkits and frameworks on explainable AI.

Toolkit	Characteristics
InterpretML[7] [270]	A python toolkit for supporting intrinsically transparent models and black-box models.
AIX360[8] [30]	A python toolkit with comprehensive set of explainability methods and datasets.
DeepExplain[9] [27]	A Tensorflow as well as Keras with Tensorflow backend toolkit for mainly supporting various gradient-based techniques and perturbation-based methods.
DIG[10] [233]	A PyTorch toolkit for supporting explainability of graph neural networks.

5.5 Future Directions

In this subsection, we discuss potential directions for future research in explainable AI. Since the interpretability of AI is a relatively new and still a developing area, many open problems need to be considered.

- **Security of explainable AI.** Recent studies have demonstrated that due to their data-driven nature, explanations of AI models are vulnerable to malicious manipulations. Attackers attempt to generate adversarial examples that can not only mislead a target classifier but also deceive its corresponding interpreter [152, 390], naturally raising potential security concerns on interpretations. Therefore, learning how to defend against adversarial attacks on explanations would be an important future direction for research.
- **Evaluation Methodologies.** Evaluation metrics are crucial for studying explanation methods; however, due to the lack of ground truths and human subjective understandings, evaluating whether the explanations are

[3]https://aix360.mybluemix.net
[4]https://github.com/interpretml/interpret
[5]https://github.com/marcoancona/DeepExplain
[6]https://github.com/divelab/DIG

reasonable and correct in regard to certain predictions is becoming intractable. The widely used evaluation methodology is based on human evaluations based on visualizing explanations, which is time-consuming and biased toward subjective human understandings. Although there are some initial studies on the evaluation of interpretability [120], it is still unclear how to measure what constitutes a good Explanation?". It is crucial to investigate qualitative and quantitative evaluations of interpretability.

- **Knowledge to Target model: from white-box to black-box.** Most existing explanation techniques require full knowledge of the explained AI system (denoted as white-box). However, knowledge regarding target AI systems is often limited in many scenarios due to privacy and security concerns. Therefore, an important direction is to understand how an explanation can be generated for making decisions in black-box systems.

6 PRIVACY

The success of modern AI systems is built upon data, and data might contain private and sensitive information – from credit card data to medical records and from social relations to family trees. To establish trustworthy AI systems, the safety of private and sensitive information carried by the data and models that could be potentially exposed throughout the AI system must be guaranteed. Therefore, increasing attention has been paid to the protection and regulation of data privacy. From a legal perspective, laws from the state level to the global level have begun to provide mandatory regulations for data privacy. For instance, the California Consumer Privacy Act (CCPA) was signed into law in 2018 to enhance privacy rights and consumer protection in California by giving consumers more control over the personal information that businesses collect; the Health Insurance Portability and Accountability Act (HIPAA) was created in 1996 to protect individual healthcare information by requiring authorization before disclosing personal healthcare information; the European Union announced General Data Protection Regulation (GDPR) to protect data privacy by giving the individual control over personal data collection and usage.

From the perspective of science and technology, although most AI technologies haven't considered privacy as the fundamental merit when they are first developed, to make modern AI systems trustworthy in privacy protection, a subfield of AI, privacy-preserving machine learning (PPML), has set privacy protection as the priority and has begun to pioneer principled approaches for preserving privacy in machine learning. Specifically, researchers uncover the vulnerabilities of existing AI systems from comprehensive studies and then develop promising technologies to mitigate these vulnerabilities. In this section, we will provide a summary of this promising and important field. Specifically, the basic concepts and taxonomy will be first discussed, and the risk of privacy breaches will be explained through various privacy attacking methods. Mainstream privacy-preserving technologies, such as *confidential computing*, *federated learning*, and *differential privacy* will be included, followed by discussions on applications in real systems, existing surveys and tools, and the future directions.

6.1 Concepts and Taxonomy

In the context of privacy protection, the adversarial goal of an attacker is to extract information about the data or machine learning models. According to the accessible information the adversary has, the attacker can be categorized into *white-box or black-box*. In a white-box setting, we assume that the attacker has all information except the data that we try to protect and the attacker aims to attack. In a black-box setting, the attacker has very limited information, for example, the query results returned by the model. Based on when the attack occurs, the privacy breach could happen in the *training phase or inference phase*. In the training phase, the adversary might be able to directly access or infer the information about the training data when she inspects or even tampers with the training process. In the inference

phase, the adversary might infer the input data of the model by inspecting the output characteristics. According to the capability of the adversary, the attacker may be *honest-but-curious or fully malicious*. An honest-but-curious attacker can inspect and monitor the training process while a fully malicious attacker can further tamper the training process. These taxonomies are not exclusive since they view the attacker from different perspectives.

6.2 Methods

We will highlight the risk of privacy leakage by introducing some representative privacy attack methods. Then, some mainstream techniques for privacy-preserving will be introduced.

6.2.1 Privacy Attack. Privacy attacks can target training data, input data, properties of data population, and even the machine learning model itself. We introduce some representative privacy attacks to reveal the risk of privacy breaches.

Membership Inference Attack. To investigate how machine learning models leak information about individual data within the training data, the membership inference attack aims to identify whether a data record is used in the training of model learning models. For instance, given the black-box access to the model, an inference model can be trained to recognize whether the given inputs are used in its training or not based on the predictions of the target model [321]. Empirically, it is shown that commonly used classification models can be vulnerable to membership inference attacks. Therefore, private information can be inferred if some user data (e.g., medical records and credit card data) is used in training the model. Please refer to the survey [179] for a comprehensive summary of membership inference attacks.

Model Inversion Attack. Model inversion attack [145, 146] aims to use the model's output to infer the information of the input data that often contain sensitive and private information. For instance, in pharmacogenetics, machine learning models are used to guide medical treatments, given the patient's genotype and demographic information. However, it has been shown that severe privacy risk exists because a patient's genetic information can be disclosed given the model and the patient's demographic information [146]. In facial recognition with neural networks, the images of people's faces can be recovered given their names, prediction confidence values, and access to the model [145]. In [392], generative adversarial networks (GANs) are used to guide the inversion process of neural networks and reconstruct high-quality facial images from face recognition classifiers. In a recent study, researchers found that the input data can be perfectly recovered through the gradient information of neural networks [401], which highlights the privacy risk in distributed learning where gradient information needs to be transmitted when people used to believe that it can preserve data privacy.

Property Inference Attack. Given the machine learning model, the property inference attack aims to extract global properties of the training dataset or training algorithm that the machine learning models do not intend to share. One example is to infer the properties that only hold for a subset of the training data or a specific class of the training data. This type of attack might leak private statistical information about the population, and the learned property can be used to exploit the vulnerability of an AI system.

Model Extraction. An adversary aims to extract the model information by querying the machine learning model in a black-box setting such that he can potentially fully reconstruct the model or create a substitute model that closely approximates the target model [341]. Once the model has been extracted, the black-box setting translates to the white-box setting, where other types of privacy attacks become much easier. Moreover, the model information typically contains an intelligent property that should be kept confidential. For instance, ML-as-a-service (MLaaS) systems, such as Amazon AWS Machine Learning, Microsoft Azure Machine Learning Studio, and Google Cloud Machine Learning

Engine, allow users to train the models on their data and provide publicly accessible query interfaces on a pay-per-query basis. The confidential model contains users' intelligent property but suffers from the risk of functionality stealing. Since the trained ML model is often considered as an intellectual property (IP), researchers recently have proposed IP protection solutions to protect ML models from infringement. The taxonomy, attacks, and evaluations in this research direction are summarized in a recent survey [367].

6.2.2 Privacy Preservation. The privacy-preserving countermeasures can be roughly categorized into three mainstream and promising directions, including confidential computing, federated learning, and differential privacy, as shown in Figure 11. Confidential computing attempts to ensure data safety during transmission and computing. Federated learning provides a new machine learning framework that allows data to be local and decentralized and avoid raw data transmission. Differential privacy aims to utilize the information about a whole dataset without exposing individual information in the dataset. Next, we review these techniques and discuss how they preserve privacy.

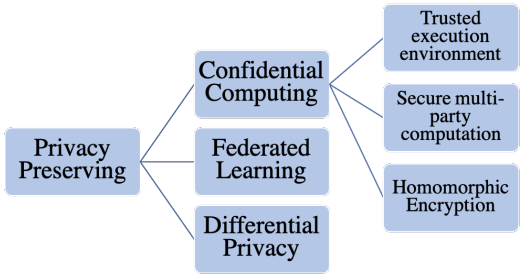


Fig. 11. An Overview of Privacy Preserving Techniques

Confidential Computing. There are mainly three types of techniques for achieving confidential computing, including Trusted Executive Environment (TEE) [304], Homomorphic Encryption (HE) [17], and Multi-party Secure Computation (MPC) [129].

Trusted Execution Environments. Trusted Execution Environments focus on developing hardware and software techniques to provide an environment that isolates data and programs from the operator system, virtual machine manager, and other privileged processes. The data is stored in the trusted execution environment (TEE) such that it is impossible to disclose or operate on the data from outside. The TEE guarantees that only authorized codes can access the protected data, and the TEE will deny the operation if the code is altered. As defined by the Confidential Computing Consortium [12], the TEE provides a level of assurance of data confidentiality, data integrity, and code integrity that essentially states that unauthorized entities cannot view, add, remove, or alter the data while it is in use within the TEE, and cannot add, remove or alter code executing in the TEE.

Secure Multi-party Computation. Secure multi-party computation (MPC) protocols aim to enable a group of data owners who might not trust one another to jointly perform a function computation that depends on all of their private input without disclosing any participant's private data. Although the concept of secure computation was primarily a theoretical interest when it was first proposed [370], it has now become a practical tool to enable privacy-preserving applications in which multiple distrusting data owners seek to compute a function cooperatively [129].

Homomorphic Encryption. Homomorphic Encryption (HE) enables computation functions on the data without accessing the plaintext by allowing mathematical operations to be performed on ciphertext without decryption. It returns the computation result in the encrypted form, which can be decrypted just as the computation is performed on the decrypted data. With partially homomorphic encryption schemes, only certain operations can be performed, which limits them to specialized problems that can be reduced as the supported operations. Fully-homomorphic encryption (FHE) schemes aim to provide support for a universal set of operations so that any finite function can be computed. The first FHE scheme was proposed by Gentry [150], and was based on lattice-based cryptography. There have been a lot of recent interests in implementing FHE schemes [90, 151], but to build a secure, deployable, scalable system using FHE is still challenging.

Federated Learning. Federated learning (FL), as shown in Figure 12, is a popular machine learning paradigm where many clients, such as mobile devices or sensors, collaboratively train machine learning models under the coordination of a central server, while keeping the training data from the clients decentralized [250]. This paradigm is in contrast with traditional machine learning settings, where the data is first collected and transmitted to the central server for further processing. In federated learning, the machine learning models are moving between the server and clients while keeping the private data locally within the clients. Therefore, it essentially avoids the transmission of private data and significantly reduces the risk of privacy breaches.

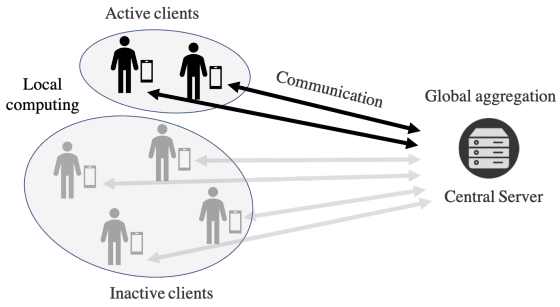


Fig. 12. Federated Learning

Next, we briefly describe a typical workflow for a federated learning system [250]:

- *Client selection*: The server samples a subset of clients from those active clients according to some eligibility requirements.
- *Broadcast*: The server broadcasts the current model and the training program to the selected clients.
- *Local computation*: The selected clients locally compute the update to the received model based on the local private data. For instance, the stochastic gradient descent (SGD) update can be run with the stochastic gradient computed based on local data and the model.
- *Aggregation*: The server collects the updated local models from the selected clients and aggregates them as an updated global model.

This workflow represents one round of the federated learning algorithm, and it will repeat until reaching specific requirements, such as convergence accuracy or performance certificates.

In addition to protecting data privacy by keeping the data local, there are many other techniques to further secure data privacy. For instance, we can apply lossy compression before transferring the models between server and clients

such that it is not easy for the adversary to infer accurate information from the model update [401]. We also can apply secure aggregation through secure multi-party computation such that no participant knows the local model information from the other participants, but the global model can still be computed [23, 261]. Additionally, we can also apply noisy perturbation to improve the differential privacy [351].

Federated learning is becoming an increasingly popular paradigm for privacy protection and has been studied, developed, and deployed in many applications. However, federated learning still faces many challenges, such as the efficiency and effectiveness of learning, especially with non-IID data distributions [203, 204, 223, 224, 234].

Differential Privacy. Differential Privacy (DP) is an area of research that aims to provide rigorous statistical guarantees for reducing the disclosure about individual information in a dataset [122, 125]. The major idea is to introduce some level of uncertainty through randomization or noise into the data such that the contribution of individual information is hidden while the algorithm can still leverage valuable information from the dataset as a whole. According to the definition [125], let's first define that the datasets D and D' are adjacent if D' can be obtained from D by altering the record of a single individual. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subset \text{Range}(\mathcal{A})$ and for all adjacent datasets D and D' such that

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\epsilon} \Pr(\mathcal{A}(D') \in \mathcal{S}) + \delta.$$

(ϵ, δ) quantifies how much information can be inferred about an individual from the output of the algorithm $\mathcal{A}$ on the dataset. For instance, if ϵ and δ are sufficiently small, the output of the algorithm will be almost identical, i.e., $\Pr[\mathcal{A}(D) \in \mathcal{S}] \approx \Pr(\mathcal{A}(D') \in \mathcal{S})$, such that it is difficult for the adversary to infer the information of any individual since the individual's contribution on the output of the algorithm is nearly masked. The privacy loss incurred by the observation ξ is defined as

$$\mathcal{L}^{\xi}_{\mathcal{A},D,D'} = \ln\left(\frac{\Pr[\mathcal{A}[D] = \xi]}{\Pr[\mathcal{A}[D'] = \xi]}\right).$$

(ϵ, δ)-differential privacy ensures that for all adjacent datasets D and D', the absolute value of the privacy loss is bounded by ϵ with probability at least $1 - \delta$. Some common methods to provide differential privacy include random response [350], Gaussian mechanism [125], Laplace mechanism [124], exponential mechanism [252], etc. Different from other techniques, differential privacy provides provable guarantees and measurements for privacy protection, but the challenge is to increase the utility under the given budget, especially for high-dimensional models.

6.3 Applications in Real Systems

Privacy-preserving techniques have been widely used to protect sensitive information in real systems. In this subsection, we discuss some representative examples.

6.3.1 Healthcare. Healthcare data can be available from patients, clinical institutions, insurance companies, pharmacies, and so on. However, the privacy concern of personal healthcare information makes it difficult to fully exploit the large-scale and diverse healthcare data to develop effective predictive models for healthcare applications. Federated learning provides an effective privacy-preserving solution for such scenarios since data across the population can be utilized while not being shared [3, 196, 291, 319, 365, 383]. Differential privacy has also gained significant attention as a general way of protecting healthcare data [111].

6.3.2 Biometric Data Analysis. Biometric data is mostly non-revocable and can be used for identification and authentication. Therefore, it is critical to protect private biometric data. To this end, confidential computing, federated learning,

and differential privacy techniques have become widely applied to protect people's biometric data such as face images, medical images, and fingerprint pattern [53, 196, 305, 347].

6.3.3 Recommender Systems. Recommender systems utilize users' interactions on products such as movies, music, and goods to provide relevant recommendations. The rating information has been shown to expose users to inference attacks, leaking private user attributes such as age, gender, etc [24, 251, 322]. To protect user privacy, recent works [202, 251, 267, 388] have developed privacy-preserving recommender systems via differential privacy.

6.3.4 Distributed Learning. In distributed learning, it is possible to recover a client's original data from their gradient information or model update [395, 401]. Secure multiparty computing is applied to protect the information of locally trained models during aggregation [23, 261, 298]. The Differentially Private SGD [15, 186, 327] provides an effective and provable manner to protect the input data by adding noise in gradient and has been popular in the training of deep learning models.

For models with a large number of trainable parameters, the differentially private models degrade the utility drastically since the added noise suffers from dimensional dependency [33]. Various subspace methods [194, 374, 375, 400] have been proposed to circumvent the dependence on the ambient dimension by leveraging the empirical observation on the low-dimensional structure of gradient space in deep networks, and they have been shown to increase the utility significantly. It is shown in [340] that transfer learning of public data from a similar domain helps improve the utility of private learning. There also exist discussions on the interplay between memorization and privacy leakage in machine learning [54, 66, 140].

6.3.5 Natural Language Processing. Recent studies [67] that can extract private data from large language models such as GTP-2 have exacerbated the privacy concerns. To mitigate these concerns, differentially private SGD (DP-SGD) has been applied to the training of large-scale training of language models such as BERT [28, 225, 373], and many techniques and tricks have been proposed to mitigate the performance drop and improve the training efficiency. It is suggested that privately learning with a pre-trained model tends to not suffer from dimension-dependent performance degradation [225, 373].

6.4 Surveys and Tools

We collect some surveys and tools relevant to privacy in AI systems for further exploration.

6.4.1 Surveys. The general concepts, threats, attack, and defense methods in privacy-preserving machine learning are summarized in several surveys [26, 112, 292]. Federated learning is comprehensively introduced in the papers [250, 368]. Differential privacy is reviewed in the surveys [122, 125, 187].

6.4.2 Tools. Popular tools and repositories in federated learning include *TensorFlow Federated (TFF)* [13], *FATE* [5], *FedML* [171], *PaddleFL* [11] and *LEAF* [6]. Popular tools in differential privacy include *Facebook Opacus* [10], *TensorFlow-Privacy* [14], *OpenDP* [9] and *Diffpriv* [299]. *Keystone Enclave* [218] is an open framework for designing Trusted Execution Environments. Popular tools in Secure Multiparty Computing and Homomorphic Encryption are summarized in the lists [7, 8]. These tools are also summarized in Table 8.

Table 8. Representative open-source toolkits and frameworks on private AI.

Toolkit	Characteristics
TensorFlow Federated [13]	An open-source framework for machine learning and other computations on decentralized data.
FedML [171]	A versatile Edge-Cloud Ecosystem for federated learning and analytics at scale.
FATE [5]	An open-source framework that supports federated learning and secure computation protocols.
LEAF [6]	A benchmarking framework for federated learning, with applications including federated learning, multi-task learning, meta-learning, and on-device learning.
Facebook Opacus [10]	A high-speed library for training PyTorch models with differential privacy.
TensorFlow-Privacy [14]	A Python library that includes implementations of TensorFlow optimizers for training machine learning models with differential privacy.
OpenDP [9]	A modular collection of statistical algorithms that adhere to the definition of differential privacy.
Diffpriv [299]	An R Package for easy differential privacy.
Keystone Enclave [218]	An open-source framework for trusted execution environments.

6.5 Future Directions

Confidential computing, federated learning, and differential privacy are three effective ways to improve privacy protection. However, they are far away from being extensively used and require more development. For instance, the computation efficiency and flexibility of confidential computing are not mature enough to support the applications of AI systems in our society. There are also great challenges to improve the efficiency and effectiveness of federated learning when deploying in large-scale and heterogeneous environments. It would be desirable to achieve a better trade-off between utility and privacy loss in differential privacy. Most importantly, a versatile and reliable system design for achieving privacy protection and different techniques should be integrated to enhance the trustworthiness of AI systems by reducing the risks of privacy leakages from multiple perspectives.

7 ACCOUNTABILITY & AUDITABILITY

In general, accountability for AI indicates how much we can trust these AI technologies and who or what we should blame if any parts of the AI technologies perform below expectation. It is about a declaration of responsibility. It is not trivial to explicitly determine the accountability for AI. On the one hand, most AI-based systems act as "black-box", due to the lack of explainability and transparency. On the other hand, real-world AI-based systems are very complex, and involve numerous key components, including input data, algorithm theory, implementation details, real-time human control, and so on. These factors further complicate the determination of accountability for AI. Although difficult and complex, it is necessary to guarantee accountability for AI. Auditability, which refers to a set of principled evaluations of the algorithm theories and implementation processes, is one of the most important methodologies in guaranteeing accountability.

It is very important to achieve a balance between accountability and innovation in AI. The overall aim is for humans to enjoy the benefits and conveniences of AI with a reliable and guarantee of safety. Additionally, however, we do not want to heavily burden the algorithm designer or put too many restrictions on end-users of AI-based systems. In this section, we discuss the accountability and auditability of AI. First, we introduce the basic concept of accountability and some key roles within it. We then describe the definition of auditability for AI and two kinds of audits. Finally, we

summarize existing surveys and tools, and discuss some future directions to enhance accountability and auditability in AI.

7.1 Concepts and Taxonomy

In this subsection, we will introduce the key concepts and taxonomies of accountability and auditability in AI.

7.1.1 Accountability. Accountability in AI has a broad definition. On the one hand, accountability can be interpreted as a property of AI. From this perspective, accountability can be improved if breakthroughs can be made in the explainability of AI algorithms. On the other hand, accountability can be referred to as a clear responsibility distribution, which focuses on who should take the responsibility for each impact of AI-based systems. Here we mainly focus on discussing the second notion. As indicated above, it is not trivial to give a clear specification for responsibility, since the operation of an AI-based system involves many different parties, such as the system designer, the system deployer, and the end-user. Any improper operation from any parties may result in system failure or potential risk. Also, all kinds of possible cases should be taken into consideration to ensure a fair distribution of responsibility. For example, the cases when an AI system does harm when working correctly versus working incorrectly should be considered differently [245, 377]. To better specify accountability, it is necessary to determine the roles and the corresponding responsibility of different parties in the function of an AI system. In [353], three roles are proposed: decision-makers, developers, and users. By refining these three roles, we propose five roles, and introduce their responsibilities and obligations as follows: **System Designers**: system designers are the designers of the AI system. They are supposed to design an AI system that meets the user requirements and is transparent and explainable to the greatest extent. It is their responsibility to offer deployment instructions and user guidelines, and to release potential risks. **Decision Makers**: decision-makers have the right to determine whether to build an AI system and what AI system should be adopted. Decision-makers should be fully aware of the benefits and risks of the candidate AI system, and take all the relevant requirements and regulations into overall consideration. **System Deployers**: system deployers are in charge of deploying an AI system. They should follow the deployment instructions carefully and ensure that the system has been deployed appropriately. Also, they are expected to offer some hands-on tutorials to the end-users. **System Auditors**: system auditors are responsible for system auditing. They are expected to provide comprehensive and objective assessments of the AI system. **End Users**: end-users are the practical operators of an AI system. They are supposed to follow the user guidelines carefully and report emerging issues to system deployers and system designers in a timely fashion.

7.1.2 Auditability. Auditability is one of the most important methodologies in ensuring accountability, which refers to a set of principled assessments from various aspects. In the IEEE standard for software development [1], an audit is defined as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures." Typically, audits can be divided into two categories as follows:

External audits: external audits [161, 308] refer to audits conducted by a third party that is independent of system designers and system deployers. External audits are expected to share no common interest with the internal workers and are likely to provide some novel perspectives for auditing the AI system. Therefore, it is expected that external audits can offer a comprehensive and objective audit report. However, there are obvious limitations to external audits. First, external audits typically cannot access all the important internal data in an AI system, such as the model training data and model implementation details [60], which increases the auditing difficulty. Additionally, external audits are

always conducted after an AI system is deployed, so that it may be costly to make adjustments over the system, and, sometimes, the system may have already done harm [264].

Internal audits: internal audits refer to audits conducted by a group of people inside the system designer or system deployer organizations. SMACTR [287] is a recent internal auditing framework proposed by researchers from Google and Partnership on AI and consists of five stages: scoping, mapping, artifact collection, testing, and reflection. Compared with external audits, internal audits can have access to a large amount of internal data, including the model training data and model implementation details, which makes internal audits much more convenient. Furthermore, internal audits can be conducted before an AI system is deployed, thus avoiding some potential harm after the system's deployment. The internal audit report can also serve as an important reference for the decision-maker to make a decision. However, an unavoidable shortcoming for internal audits is that they share the same interest as the audited party, which makes it challenging to give an objective audit report.

7.2 Surveys and Tools

In this subsection, we summarize existing surveys and tools about accountability and auditability of AI, to facilitate readers who want to explore this field further.

7.2.1 Surveys. A recent work on algorithmic accountability is presented in [353]. It takes Boven's definition of accountability [52] as the basic concept and combines it with numerous literature in algorithmic accountability to build the concept's definition.

7.2.2 Tools. The other five dimensions (safety & robustness, non-discrimination & fairness, explainability, privacy, environmental well-being) discussed in this survey are also important aspects to be evaluated during algorithm auditing. Therefore, most tools introduced in section 3.6, 4.4.2, 5.4.2, 6.4, and 8.3 can also be used for the purpose of auditing.

7.3 Future Directions

For accountability, it is important to further enhance the explainability of the AI system. Only when we have a deep and thorough understanding of its theory and mechanism can we fully rely on it or make a well-recognized responsibility distribution scheme. For auditability, it is always a good option to conduct both external audits and internal audits, so that we can have a comprehensive and objective overview of an AI system. Furthermore, we need to be aware that an AI system is constantly dynamic. It can change with input data and environment. Thus, to make an effective and timely audit, it is necessary to audit the system periodically and to update auditing principles with the system changes [228].

8 ENVIRONMENTAL WELL-BEING

A trustworthy AI system should be sustainable and environmentally friendly [325]. In fact, the large-scale development and deployment of AI systems bring a huge burden of energy consumption, which inevitably affects the environment. For example, Table 9 shows the carbon emission (as an indicator of energy consumption) of training NLP models and that of daily consumption [330]. We find that training a common NLP pipeline has the same carbon emissions as a human produces in seven years. Training and fine-tuning a large Transformer model costs five times more energy consumption than a car over its lifetime. Besides model development, in other areas, such as data center cooling,[11] there is also a huge energy cost. The rapid development of AI technology further challenges the tense global situation

[11] https://www.forbes.com/sites/forbestechcouncil/2020/08/17/why-we-should-care-about-the-environmental-impact-of-ai/?sh=b90512e56ee2

of energy shortage and environmental deterioration. Hence, environmental friendliness becomes an important issue to consider in building trustworthy AI systems. In this section, we review the existing works regarding the environmental impacts of AI technologies. Existing works mainly focus on the impact of AI systems' energy consumption on the environment. We first present an overview of the strategies for reducing energy consumption, e.g., model compression, and then introduce the works on estimating the energy consumption and evaluating the environmental impacts of real-world AI systems in different domains. Finally, we summarize the existing surveys and tools on this dimension.

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 9. Comparsions between estimated CO_2 emissions produced by daily lives and training NLP models. (Table Credit: [330])

8.1 Methods

In this subsection, we summarize the techniques developed for reducing the energy use of AI algorithms. Improving the energy efficiency of AI systems involves algorithm-level and hardware-level solutions. We will introduce two common classes of algorithm-level approaches: model compression and adaptive design, as well as hardware-level energy-saving methods.

8.1.1 Model Compression. Model compression is a hot topic in deep learning that receives continuous attention from both academia and industry. It studies how to reduce the size of a deep model to save storage space and the energy consumption for training and deploying models, with an acceptable sacrifice on model performance. For CNN models in the image domain, parameter pruning and quantization [92, 167], low-rank factorization [185, 293], transferred/compact convolutional filters [97, 318], and knowledge distillation have been proposed [176, 296]. Similarly, in the text domain, researchers borrow and extend these methods: pruning [65, 256], quantization [88, 177], knowledge distillation [207, 331, 334], and parameter sharing [113, 216], to compress popular NLP models, such as Transformer and BERT.

8.1.2 Adaptive Design. Another line of research focuses on adaptively designing a model architecture to optimize the energy efficiency of a model. Yang et al. [369] propose a pruning approach to design CNN architectures to achieve an energy-saving goal. In their method, the model is pruned in a layer-by-layer manner, where the layer that consumes the most energy is pruned first. Stamoulis et al. [328] propose a framework to adaptively design CNN models for image classification under energy consumption restrictions. They formulate the design of a CNN architecture as a hyperparameter optimization problem and solve it by Bayesian optimization.

8.1.3 Hardware. In addition to the algorithm level, endeavors are also conducted to improve the energy efficiency of AI from the design of the hardware. Computing devices or platforms specially designed for AI applications are proposed to maximize the training and inference efficiency of AI algorithms. Specifically, hardware designed for DNN models are called DNN accelerators [85]. Esmaeilzadeh et al. [127] design a neural processing unit (NPU) to execute the fixed computations of a neuron such as multiplication, accumulation, and sigmoid, on chips. Later, Liu et al. [235] proposed RENO, which is a more advanced on-chip architecture for neural network acceleration. There is also hardware designed for specific NN structures. Han et al. [166] investigate how to design an efficient computation device for a sparse neural network, where weight matrices and feature maps are sparse. They [165] also devise an efficient speech recognition engine that is dedicated to RNN models. Furthermore, ReGAN [77] is developed to accelerate generative adversarial networks (GANs).

8.2 Applications in Real Systems

As described before, the environmental impacts of AI systems mainly come from energy consumption. In this subsection, we introduce the research on evaluating and estimating the energy consumption of real-world AI systems in different domains.

In the field of computer vision, Li et al. [222] first investigate the energy use of CNNs on image classification tasks. They provide a detailed comparison among different types of CNN layers, and also analyze the impact of hardware on energy consumption. Cai et al. [61] introduce the framework NeuralPower which can estimate the power and runtime across different layers in a CNN, to help developers to understand the energy efficiency of their models before deployment. They also propose evaluating CNN models with a novel metric "energy-precision ratio". Based on it, developers can trade off energy consumption and model performance according to their own needs, and choose the appropriate CNN architecture. In the field of NLP, Strubell et al. [330] examine the carbon emissions of training popular NLP models, namely, Transformer, ELMo, and GPT-2, on different types of hardware, and shed light on the potential environmental impacts of NLP research and applications.

8.3 Surveys and Tools

In this subsection, we collect related surveys and tools on the dimension of environmental well-being.

8.3.1 Surveys. From the algorithm-level perspective, García-Martín et al. [147] present a comprehensive survey on energy consumption estimation methods from both the computer architecture and machine learning communities. Mainly, they provide a taxonomy for the works in computer architecture and analyze the strengths and weaknesses of the methods in various categories. Cheng et al. [87] summarize the common model compression techniques and organize them into four categories, and then present a detailed analysis on the performance, application scenarios, advantages, and disadvantages of each category. In the hardware-level perspective, Wang et al. [348] compares the performance and energy consumption of the processors from different vendors for AI training. Mittal and Vetter [259] review the approaches for analyzing and improving GPU energy efficiency. The survey [85] summarizes the latest progress on DNN accelerator design.

8.3.2 Tools. *SyNERGY* [295] is a framework integrated with Caffe for measuring and predicting the energy consumption of CNNs. Lacoste et al. [214] develop a *Machine Learning Emissions Calculator* as a tool to quantitatively estimate the carbon emissions of training an ML model, which can enable researchers and practitioners to better understand the environmental impact caused by their models. *Accelergy* [361] and *Timeloop* [276] are two representative energy

estimation tools for DNN accelerators. A summary of representative open-source tooklits on environmental well-being in AI can be found in Table 10.

Table 10. Representative open-source toolkits and frameworks on environmental well-being in AI.

Toolkit	Characteristics
SyNERGY [295]	A Caffe-based framework for CNN energy consumption estimation.
MLEC [214]	A tool for estimating the carbon emissions of training a ML model.
Accelergy [361]	An architecture-level energy estimation tool for DNN accelrators.
Timeloop [276]	A systematic DNN accelerator evaluation tool that supports energy estimation.

8.4 Future Directions

Research on reducing the energy consumption of AI systems for environmental well-being is on the rise. At the algorithmic level, automated machine learning (AutoML), which aims to automatically design effective and efficient model architectures for certain tasks, emerges as a novel direction in the AI community. Existing works in AutoML focus more on designing an algorithm to improve its performance, but don't usually treat energy consumption savings as the highest priority. Using AutoML technologies to design energy-saving models needs further exploration in the future. At the hardware level, current research on DNN accelerators pays more attention to devising efficient deployment devices to facilitate model inference, but the procedure of model training is overlooked. The design of efficient customized training devices for various DNN models is a practical and promising direction to investigate in the future.

9 INTERACTIONS AMONG DIFFERENT DIMENSIONS

An ideal trustworthy AI system should simultaneously satisfy the six dimensions discussed above. In reality, the six dimensions are not independent of one another. The satisfaction of one dimension can promote the pursuit of another dimension. Meanwhile, conflicts exist among different dimensions. The realization of one dimension could violate another dimension, which makes it impossible for two or more dimensions to be met simultaneously in some scenarios. Researchers and practitioners should be aware of the complicated interactions among different dimensions. Knowing the accordance between two dimensions brings us an alternative idea to achieve one dimension: we can try to satisfy one dimension by realizing the other. Moreover, when two dimensions are contradictory, we can make a trade-off between them according to our needs. In this section, we discuss some known accordance and conflict interactions among different dimensions.

9.1 Accordance

Two dimensions are accordant when the satisfaction of one dimension can facilitate the achievement of the other, or the two dimensions promote each other. Next, we show two examples of accordance interactions among dimensions.

Robustness & Explainability. Studies show that deep learning models' robustness against adversarial attacks positively correlates with their explainability [128, 133, 269]. Etmann et al. [128] find that models trained with robustness objectives show more interpretable saliency maps. Specifically, they prove rigorously in mathematics that Lipschitz regularization, which is commonly used for robust training, forces the gradients to align with the inputs. Noack et al. [269] further investigate the opposite problem: will an interpretable model be more robust? They propose Interpretation Regularization (IR) to train models with explainable gradients and empirically show that a model can be more robust to adversarial attacks if it is trained to produce explainable gradients.

Fairness & Environmental Well-being. Fairness in the field of AI is a broad topic, which involves not only the fairness of AI service providers and users, but also the equality of AI researchers. As mentioned in section 8, the development trend of deep learning models toward larger models and more computing resource consumption not only causes adverse environmental impact but also aggravates the inequality of research [330], since most researchers cannot afford high-performance computing devices. Hence, the efforts for ensuring the environmental well-being of AI techniques, such as reducing the cost of training large AI models, are in accordance with the fairness principle of trustworthy AI.

9.2 Conflict

Two dimensions are conflicting when the satisfaction of one dimension hinders the realization of the other. Next, we show three examples of the conflicting interactions among dimensions.

Robustness & Privacy. Recent studies find tensions between the robustness and the privacy requirements of trustworthy AI. Song et al. [326] check how the robust training against adversarial attacks influences the risk of a model against membership inference attack. They find that models trained with adversarial defense approaches are more likely to expose sensitive information in training data via membership inference attacks. The reason behind this is that models trained to be robust to adversarial examples typically overfit to training data, which makes training data easier to be detected from models' outputs. From the perspective of the victims of unauthorized AI applications, the conflict between robustness and privacy is not always bad. Many large-scale AI applications, such as face recognition systems, are trained on data collected from people without their explicit consent, which raises concerns about the disclosure of personal information [280]. Thus, recent studies propose to take advantage of the defects in robustness of unauthorized AI models to prevent them from exploiting private data [180, 317, 335]. Specifically, Shan et al. [317] introduce imperceptible noise into users' photos, so that it is hard for anyone who collects such images to train a face recognition model to correctly identify the users. Huang et al. [180] propose to add an error-minimizing noise into private training examples to fool the deep learning models into "believing there is nothing to learn from these examples". It is worth mentioning that such poisoning attacks cannot solve the privacy issue once and for all. Models trained adaptively against the perturbed examples or new technologies developed after the attacks still have potential for nullifying the attacks [285].

Robustness & Fairness. Robustness and fairness can also conflict with each other in particular scenarios. As discussed in section 3, adversarial training is one of the mainstream approaches for improving the robustness of a deep learning model. Recent research [363] indicates that adversarial training can introduce a significant disparity of performance and robustness among different groups, even if the datasets are balanced. Thus, the adversarial training algorithm improves the robustness of a model at the expense of its fairness. Accordingly, the work [363] proposes a framework called Fair-Robust-Learning (FRL) to ensure fairness while improving a model's robustness.

Fairness & Privacy. Cummings et al. [105] investigate the compatibility of fairness and privacy of classification models, and theoretically prove that differential privacy and exact fairness in terms of equal opportunity are unlikely to be achieved simultaneously. By relaxing the condition, this work further shows that it is possible to find a classifier that satisfies both differential privacy and approximate fairness.

Addressing the Conflicts among Dimensions. As discussed in this subsection, different dimensions of trustworthy AI can interact with one another in a conflicting manner. The conflicts make different dimensions restrict each other so sometimes we cannot meet them simultaneously to build a completely trustworthy AI system. Thus, how to deal with the conflicts among dimensions becomes a future research direction that draws increasing attention. There are two

potential directions. First, a direct problem to be solved is how to alleviate the conflict so as to meet both sides to the greatest extent [363]. Second, if the tension is proven to be unsolvable, more research on this field is needed for us to better understand how one dimension impacts one another, and how we can sacrifice one for another based on our actual needs [105].

10 FUTURE DIRECTIONS

In this survey, we elaborate on six of the most concerning and crucial dimensions an AI system needs to meet to be trustworthy. Beyond that, some dimensions have not received extensive attention, but are worth exploring in the future. In this section, we will discuss several other potential dimensions of trustworthy AI.

10.1 Human agency and oversight

The ethical guidelines for trustworthy AI proposed by different countries and regions all emphasize the human autonomy principle of AI technology [325]. Human autonomy prohibits AI agents from subordinating, coercing, or manipulating humans, and requires humans to maintain self-determination over themselves. To achieve the principle of human autonomy, the design of AI systems should be human-centered. Specifically, human agency and oversight should be guaranteed in the development and deployment of AI systems. Human agency enables humans to make decisions independently based on the outputs of an AI system, instead of being totally subject to AI's decisions. A desirable human agency technology encourages users to understand the mechanism of an AI system and enables users to evaluate and challenge the decisions of an AI system, and make better choices by themselves. Human oversight enables humans to oversee AI systems throughout their life cycle, from design to usage. It can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command governance strategies.

10.2 Creditability

With the wide deployment of AI systems, people increasingly rely on content produced or screened by AI, such as an answer to a question given by a question-answering (QA) agent or a piece of news delivered by a recommender system. However, the integrity of such content is not always guaranteed. For example, an AI system that exposes users to misinformation should not be considered trustworthy. Hence, additional mechanisms and approaches should be incorporated in AI systems to ensure their creditability.

10.3 Interactions among Different Dimensions

The research on the interactions among different dimensions is still in an early stage. Besides the several instances shown in this paper, there are potential interactions between other dimension pairs remaining to be investigated. For example, people may be interested in the relationship between fairness and interpretability. In addition, the interaction formed between two dimensions can be different in different scenarios, which needs further exploration. For example, an interpretable model may promote its fairness by making its decision process transparent. On the contrary, techniques to improve the interpretability of a model may introduce a disparity of interpretability among different groups, which leads to a fairness problem. Although there are numerous problems to study, understanding the interactions among different dimensions is very important in building a trustworthy AI system.

11 CONCLUSION

In this survey, we present a comprehensive overview of trustworthy AI from a computational perspective and clarify the definition of trustworthy AI from multiple perspectives, distinguishing it from similar concepts. We introduce six of the most crucial dimensions that make an AI system trustworthy; namely, Safety & Robustness, Nondiscrimination & Fairness, Explainability, Accountability & Auditability, Privacy, and Environmental Well-being. For each dimension, we present an overview of related concepts and a taxonomy to help readers understand, how each dimension is studied, and summarize the representative technologies, to enable readers to follow the latest research progress in each dimension. To further deepen the understanding of each dimension, we provide numerous examples of applications in real-world systems and summarize existing related surveys and tools. We also discuss potential future research directions within each dimension. We then analyze the accordance and conflicting interactions among different dimensions. Finally, it is important to mention that outside of the six dimensions elaborated in this survey, there still exist some other potential issues that may undermine our trust in AI systems. Hence, we discuss several possible dimensions of trustworthy AI as future research directions.

REFERENCES

[1] 2008. 1028-2008 - IEEE Standard for Software Reviews and Audits. In *1028-2008 - IEEE Standard for Software Reviews and Audits*.

[2] 2017. The Montreal Declaration of Responsible AI. https://www.montrealdeclaration-responsibleai.com/the-declaration Accessed March 18, 2021.

[3] 2018. Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics* 112 (2018), 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

[4] 2019. Governance Principles for the New Generation Artificial Intelligence–Developing Responsible Artificial Intelligence. https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html Accessed March 18, 2021.

[5] 2021. Federated AI Technology Enabler. https://fate.fedai.org/.

[6] 2021. LEAF: A Benchmark for Federated Settings. https://leaf.cmu.edu/.

[7] 2021. A list of Homomorphic Encryption libraries, software or resources. https://github.com/jonaschn/awesome-he.

[8] 2021. A list of MPC software or resources. https://github.com/rdragos/awesome-mpc.

[9] 2021. OenDP: Open Source Tools for Differential Privacy. https://opendp.org/.

[10] 2021. Opacus: Train PyTorch models with Differential Privacy. https://opacus.ai/.

[11] 2021. Paddle Federated Learning. https://github.com/PaddlePaddle/PaddleFL.

[12] 2021. A Technical Analysis of Confidential Computing. https://confidentialcomputing.io/wp-content/uploads/sites/85/2021/03/CCC-Tech-Analysis-Confidential-Computing-V1.pdf Accessed Jan, 2021.

[13] 2021. TensorFlow Federated. https://github.com/tensorflow/federated.

[14] 2021. TensorFlow Privacy. https://github.com/tensorflow/privacy.

[15] 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 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[16] Talal AA Abdullah, Mohd Soperi Mohd Zahid, and Waleed Ali. 2021. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry* 13, 12 (2021), 2439.

[17] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–35.

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

[19] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

[20] 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. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[21] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[22] 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.

[23] 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 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1231–1247.

[24] Esma Aïmeur, Gilles Brassard, José M Fernandez, and Flavien Serge Mani Onana. 2008. A lambic: a privacy-preserving recommender system for electronic commerce. *International Journal of Information Security* 7, 5 (2008), 307–334.

[25] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access* 6 (2018), 14410–14430.

[26] Mohammad Al-Rubaie and J Morris Chang. 2019. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy* 17, 2 (2019), 49–58.

[27] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*.

[28] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-Scale Differentially Private BERT. arXiv:2108.01624 [cs.LG]

[29] 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. *Information Fusion* 58 (2020), 82–115.

[30] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilovic, et al. 2020. Ai explainability 360: An extensible toolkit for understanding data and machine learning models. *Journal of Machine Learning Research* 21, 130 (2020), 1–6.

[31] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *International Conference on Machine Learning*. PMLR, 405–413.

[32] Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. 2010. The security of machine learning. *Machine Learning* 81, 2 (2010), 121–148.

[33] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 464–473.

[34] Christine Raouf Saad Basta, Marta Ruiz Costa-Jussà, and José Adrián Rodríguez Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. Association for Computational Linguistics, 99–102.

[35] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. 2019. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27.

[36] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[37] Vaishak Belle and Ioannis Papantonis. 2020. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698* (2020).

[38] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[39] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[40] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 387–402.

[41] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012).

[42] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

[43] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[44] Aleksandar Bojcheski and Stephan Günnemann. 2018. Adversarial attacks on node embeddings. *arXiv preprint arXiv:1809.01093* (2018).

[45] Aleksandar Bojchevski and Stephan Günnemann. 2019. Adversarial Attacks on Node Embeddings via Graph Poisoning. arXiv:1809.01093 [cs.LG]

[46] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520* (2016).

[47] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[48] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035* (2019).

[49] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035* (2019).

[50] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*. 491–500.

[51] Avishek Joey Bose and William L Hamilton. 2019. Compositional fairness constraints for graph embeddings. *arXiv preprint arXiv:1905.10674* (2019).

[52] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal* 13, 4 (2007), 447–468.

[53] Julien Bringer, Hervé Chabanne, and Alain Patey. 2013. Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends. *IEEE Signal Processing Magazine* 30, 2 (2013), 42–52.

[54] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. 2021. *When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?* Association for Computing Machinery, New York, NY, USA, 123–132. https://doi.org/10.1145/3406325.3451131

[55] 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 preprint arXiv:2004.07213* (2020).

[56] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*. PMLR, 803–811.

[57] Bruce G Buchanan. 2005. A (very) brief history of artificial intelligence. *Ai Magazine* 26, 4 (2005), 53–53.

[58] Jacques Bughin, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. 2018. Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute* (2018).

[59] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[60] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[61] Ermao Cai, Da-Cheng Juan, Dimitrios Stamoulis, and Diana Marculescu. 2017. Neuralpower: Predict and deploy energy-efficient convolutional neural networks. In *Asian Conference on Machine Learning*. PMLR, 622–637.

[62] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[63] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[64] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*. Springer, 43–57.

[65] Shijie Cao, Chen Zhang, Zhuliang Yao, Wencong Xiao, Lanshun Nie, Dechen Zhan, Yunxin Liu, Ming Wu, and Lintao Zhang. 2019. Efficient and effective sparse LSTM on FPGA with bank-balanced sparsity. In *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. 63–72.

[66] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 267–284.

[67] 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 *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[68] 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.

[69] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. arXiv:1608.04644 [cs.CR]

[70] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 1–7.

[71] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. 2019. Unlabeled Data Improves Adversarial Robustness. arXiv:1905.13736 [stat.ML]

[72] S Carty. 2011. Many Cars Tone Deaf To Women's Voices. *AOL Autos* (2011).

[73] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. *arXiv preprint arXiv:2010.04053* (2020).

[74] L. Celis, Amit Deshpande, Tarun Kathuria, and N. Vishnoi. 2016. How to be Fair and Diverse? *ArXiv* abs/1610.07183 (2016).

[75] L Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443* (2019).

[76] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[77] Fan Chen, Linghao Song, and Yiran Chen. 2018. Regan: A pipelined reram-based accelerator for generative adversarial networks. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 178–183.

[78] Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. 2019. Robust decision trees against adversarial examples. In *International Conference on Machine Learning*. PMLR, 1122–1131.

[79] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* (2018).

[80] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[81] Jinyin Chen, Yangyang Wu, Xuanheng Xu, Yixian Chen, Haibin Zheng, and Qi Xuan. 2018. Fast Gradient Attack on Network Embedding. arXiv:1809.02797 [physics.soc-ph]

[82] Steven Chen, Nicholas Carlini, and David Wagner. 2020. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*. 30–39.

[83] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally fair clustering. In *International Conference on Machine Learning*. PMLR, 1032–1041.

[84] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv:1712.05526 [cs.CR]

[85] Yiran Chen, Yuan Xie, Linghao Song, Fan Chen, and Tianqi Tang. 2020. A survey of accelerator architectures for deep neural networks. *Engineering* 6, 3 (2020), 264–274.

[86] Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. 2018. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *arXiv preprint arXiv:1803.01128* (2018).

[87] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. 2017. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282* (2017).

[88] Robin Cheong and Robel Daniel. 2019. *transformers. zip: Compressing Transformers with Pruning and Quantization.* Technical Report. Technical report, Stanford University, Stanford, California.

[89] S. Chiappa. 2019. Path-Specific Counterfactual Fairness. In *AAAI*.

[90] Ilaria Chillotti, Nicolas Gama, Mariya Georgieva, and Malika Izabachene. 2016. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In *international conference on the theory and application of cryptology and information security*. Springer, 3–33.

[91] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 173–181.

[92] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. 2016. Towards the limit of network quantization. *arXiv preprint arXiv:1612.01543* (2016).

[93] A. Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5 2 (2017), 153–163.

[94] Alexandra Chouldechova and Max G'Sell. 2017. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046* (2017).

[95] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 1310–1320.

[96] Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy!. In *Seventh international AAAI conference on weblogs and social media*.

[97] Taco Cohen and Max Welling. 2016. Group equivariant convolutional networks. In *International conference on machine learning*. PMLR, 2990–2999.

[98] EC HLEG AI-European Commission et al. 2019. Independent High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI* (2019).

[99] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[100] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[101] Bo Cowgill and Catherine Tucker. 2017. Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms* (2017).

[102] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: a standardized adversarial robustness benchmark. arXiv:2010.09670 [cs.LG]

[103] Francesco Croce and Matthias Hein. 2020. Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. arXiv:1907.02044 [cs.LG]

[104] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. arXiv:2003.01690 [cs.LG]

[105] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 309–315.

[106] Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. 72–78.

[107] Enyan Dai and Suhang Wang. 2021. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 680–688.

[108] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial Attack on Graph Structured Data. arXiv:1806.02371 [cs.LG]

[109] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 99–108.

[110] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711* (2020).

[111] Fida Kamal Dankar and Khaled El Emam. 2013. Practicing differential privacy in health care: A review. *Trans. Data Priv.* 6, 1 (2013), 35–67.

[112] Emiliano De Cristofaro. 2020. An overview of privacy in machine learning. *arXiv preprint arXiv:2005.08679* (2020).

[113] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819* (2018).

[114] Li Deng and Yang Liu. 2018. *Deep learning in natural language processing*. Springer.

[115] Matthew James Denny and Arthur Spirling. 2016. Assessing the consequences of text preprocessing decisions. *Available at SSRN* (2016).

[116] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[117] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens Are Powerful Too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8173–8188.

[118] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. 2019. AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch. *arXiv preprint arXiv:1902.07623* (2019).

[119] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[120] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[121] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[122] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[123] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[124] 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.

[125] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.

[126] Danielle Ensign, Sorelle A Friedler, Scott Nevlle, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. In *Proceedings of Algorithmic Learning Theory,*, Vol. 83.

[127] Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger. 2012. Neural acceleration for general-purpose approximate programs. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 449–460.

[128] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. 2019. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172* (2019).

[129] D. Evans, V. Kolesnikov, and M. Rosulek. 2018. . https://doi.org/10.1561/3300000019

[130] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2017. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945* (2017).

[131] Wenqi Fan, Tyler Derr, Yao Ma, Jianping Wang, Jiliang Tang, and Qing Li. 2019. Deep Adversarial Social Recommendation. In *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*. 1351–1357.

[132] Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li. 2021. Attacking Black-box Recommendations via Copying Cross-domain User Profiles. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 1583–1594.

[133] Wenqi Fan, Wei Jin, Xiaorui Liu, Han Xu, Xianfeng Tang, Suhang Wang, Qing Li, Jiliang Tang, Jianping Wang, and Charu Aggarwal. 2021. Jointly Attacking Graph Neural Network and its Explanations. *arXiv preprint arXiv:2108.03388* (2021).

[134] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The World Wide Web Conference*. 417–426.

[135] Wenqi Fan, Yao Ma, Qing Li, Jianping Wang, Guoyong Cai, Jiliang Tang, and Dawei Yin. 2020. A Graph Neural Network Framework for Social Recommendations. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[136] Wenqi Fan, Yao Ma, Dawei Yin, Jianping Wang, Jiliang Tang, and Qing Li. 2019. Deep social collaborative filtering. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 305–313.

[137] Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. 2018. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th Annual Computer Security Applications Conference*. 381–392.

[138] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015).

[139] 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.

[140] Vitaly Feldman. 2020. *Does Learning Require Memorization? A Short Tale about a Long Tail*. Association for Computing Machinery, New York, NY, USA, 954–959. https://doi.org/10.1145/3357713.3384290

[141] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341* (2019).

[142] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (2019), 1287–1289.

[143] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28, 4 (2018), 689–707.

[144] World Economic Forum. 2020. The Future of Jobs Report 2020. World Economic Forum, Geneva, Switzerland.

[145] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.

[146] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 17–32.

[147] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. 2019. Estimation of energy consumption in machine learning. *J. Parallel and Distrib. Comput.* 134 (2019), 75–88.

[148] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3356–3369.

[149] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[150] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 169–178.

[151] Craig Gentry and Shai Halevi. 2011. Implementing gentry's fully-homomorphic encryption scheme. In *Annual international conference on the theory and applications of cryptographic techniques*. Springer, 129–148.

[152] 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.

[153] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[154] Naman Goel, Mohammad Yaghini, and B. Faltings. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018).

[155] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[156] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[157] Hila Gonen and Kellie Webster. 2020. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1991–1995.

[158] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. 2017. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960* (2017).

[159] S.T. Gooden. 2015. *Race and Social Equity: A Nervous Area of Government*. Taylor & Francis. https://books.google.com/books?id=y2dsBgAAQBAJ

[160] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[161] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

[162] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. 2017. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* (2017).

[163] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[164] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

[165] Song Han, Junlong Kang, Huizi Mao, Yiming Hu, Xin Li, Yubin Li, Dongliang Xie, Hong Luo, Song Yao, Yu Wang, et al. 2017. Ese: Efficient speech recognition engine with sparse lstm on fpga. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. 75–84.

[166] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A Horowitz, and William J Dally. 2016. EIE: Efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 243–254.

[167] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[168] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).

[169] Moritz Hardt, E. Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

[170] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[171] Chaoyang He, Songze Li, Jinhyun So, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, Li Shen, Peilin Zhao, Yan Kang, Yang Liu, Ramesh Raskar, Qiang Yang, Murali Annavaram, and Salman Avestimehr. 2020. FedML: A Research Library and Benchmark for Federated Machine Learning. *arXiv preprint arXiv:2007.13518* (2020).

[172] Úrsula Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. 2017. Calibration for the (Computationally-Identifiable) Masses. *ArXiv* abs/1711.08513 (2017).

[173] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.

[174] Katherine L Hermann and Andrew K Lampinen. 2020. What shapes feature representations? exploring datasets, architectures, and training. *arXiv preprint arXiv:2006.12433* (2020).

[175] Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law* 20, 1 (2019), 83–121.

[176] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[177] Lu Hou and James T Kwok. 2018. Loss-aware weight quantization of deep networks. *arXiv preprint arXiv:1802.08635* (2018).

[178] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics* 24, 5 (2018), 1521–1536.

[179] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. 2021. Membership Inference Attacks on Machine Learning: A Survey. arXiv:2103.07853 [cs.LG]

[180] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2020. Unlearnable Examples: Making Personal Data Unexploitable. In *International Conference on Learning Representations*.

[181] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2021. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898* (2021).

[182] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361* (2020).

[183] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[184] Vasileios Iosifidis, B. Fetahu, and Eirini Ntoutsi. 2019. FAE: A Fairness-Aware Ensemble Framework. *2019 IEEE International Conference on Big Data (Big Data)* (2019), 1375–1380.

[185] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866* (2014).

[186] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing Differentially Private Machine Learning: How Private is Private SGD?. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 22205–22216. https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf

[187] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. 2014. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584* (2014).

[188] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 702–712.

[189] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. 2020. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* 2, 10 (2020), 573–584.

[190] Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, and Jiliang Tang. 2020. Adversarial attacks and defenses on graphs: A review and empirical study. *arXiv preprint arXiv:2003.00653* (2020).

[191] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. 2020. Graph Structure Learning for Robust Graph Neural Networks. arXiv:2005.10203 [cs.LG]

[192] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. 2017. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309* (2017).

[193] Matthew Joseph, M. Kearns, Jamie H. Morgenstern, Seth Neel, and A. Roth. 2016. Fair Algorithms for Infinite and Contextual Bandits. *arXiv: Learning* (2016).

[194] Peter Kairouz, Monica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. 2021. (Nearly) Dimension Independent Private ERM with AdaGrad Rates via Publicly Estimated Subspaces. In *Proceedings of Thirty Fourth Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 134)*, Mikhail Belkin and Samory Kpotufe (Eds.). PMLR, 2717–2746. https://proceedings.mlr.press/v134/kairouz21a.html

[195] Peter Kairouz, Jiachun Liao, Chong Huang, and Lalitha Sankar. 2019. Censored and fair universal representations using generative adversarial models. *arXiv preprint arXiv:1910.00411* (2019).

[196] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2, 6 (2020), 305–311.

[197] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[198] F. Kamiran and T. Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2011), 1–33.

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

[200] F. Kamiran and I. Žliobaitė. 2013. Explainable and Non-explainable Discrimination in Classification. In *Discrimination and Privacy in the Information Society*.

[201] Toshihiro Kamishima, S. Akaho, Hideki Asoh, and J. Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML/PKDD*.

[202] Michael Kapralov and Kunal Talwar. 2013. On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1395–1414.

[203] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*. PMLR, 5132–5143.

[204] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. 2020. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4519–4529.

[205] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, D. Janzing, and B. Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *NIPS*.

[206] M. P. Kim, O. Reingold, and G. N. Rothblum. 2018. Fairness Through Computationally-Bounded Awareness. In *NeurIPS*.

[207] Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947* (2016).

[208] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[209] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[210] Joost N Kok, Egbert J Boers, Walter A Kosters, Peter Van der Putten, and Mannes Poel. 2009. Artificial intelligence: definition, trends, techniques, and cases. *Artificial intelligence* 1 (2009), 270–299.

[211] 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 2018 World Wide Web Conference*. 853–862.

[212] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[213] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*.

[214] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700* (2019).

[215] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[216] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[217] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 1995.

[218] Dayeol Lee, David Kohlbrenner, Shweta Shinde, Krste Asanovic, and Dawn Song. 2020. Keystone: An Open Framework for Architecting Trusted Execution Environments. In *Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys '20)*.

[219] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[220] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, and Michael Witbrock. 2018. Discrete Attacks and Submodular Optimization with Applications to Text Classification. *CoRR* abs/1812.00151 (2018). arXiv:1812.00151 http://arxiv.org/abs/1812.00151

[221] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. 2011. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 163–168.

[222] Da Li, Xinbo Chen, Michela Becchi, and Ziliang Zong. 2016. Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)*. IEEE, 477–484.

[223] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*, Inderjit S. Dhillon, Dimitris S. Papailiopoulos, and Vivienne Sze (Eds.). mlsys.org. https://proceedings.mlsys.org/book/316.pdf

[224] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2020. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJxNAnVtDS

[225] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2021. Large Language Models Can Be Strong Differentially Private Learners. arXiv:2110.05679 [cs.LG]

[226] Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. 2020. DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses. arXiv:2005.06149 [cs.LG]

[227] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2021), 18.

[228] Sebastian Lins, Stephan Schneider, Jakub Szefer, Shafeeq Ibraheem, and Ali Sunyaev. 2019. Designing monitoring systems for continuous certification of cloud services: deriving meta-requirements and design guidelines. *Communications of the Association for Information Systems* 44, 1 (2019), 25.

[229] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does Gender Matter? Towards Fairness in Dialogue Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4403–4416.

[230] Haochen Liu, Tyler Derr, Zitao Liu, and Jiliang Tang. 2019. Say What I Want: Towards the Dark Side of Neural Dialogue Models. *arXiv preprint arXiv:1909.06044* (2019).

[231] Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The Authors Matter: Understanding and Mitigating Implicit Bias in Deep Text Classification. *arXiv preprint arXiv:2105.02778* (2021).

[232] Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating Gender Bias for Neural Dialogue Generation with Adversarial Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 893–903.

[233] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, et al. 2021. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. *arXiv preprint arXiv:2103.12608* (2021).

[234] Xiaorui Liu, Yao Li, Rongrong Wang, Jiliang Tang, and Ming Yan. 2021. Linear Convergent Decentralized Optimization with Compression. In *International Conference on Learning Representations*. https://openreview.net/forum?id=84gjULz1t5

[235] Xiaoxiao Liu, Mengjie Mao, Beiye Liu, Hai Li, Yiran Chen, Boxun Li, Yu Wang, Hao Jiang, Mark Barnell, Qing Wu, et al. 2015. RENO: A high-efficient reconfigurable neuromorphic computing accelerator design. In *Proceedings of the 52nd Annual Design Automation Conference*. 1–6.

[236] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[237] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2016. Learning to pivot with adversarial networks. *arXiv preprint arXiv:1611.01046* (2016).

[238] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, 189–202.

[239] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017), 4765–4774.

[240] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573* (2020).

[241] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In *The World Wide Web Conference*. 1210–1221.

[242] Yao Ma, Suhang Wang, Tyler Derr, Lingfei Wu, and Jiliang Tang. 2019. Attacking Graph Convolutional Networks via Rewiring. arXiv:1906.03750 [cs.LG]

[243] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[244] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2012. Collaborative filtering and the missing at random assumption. *arXiv preprint arXiv:1206.5267* (2012).

[245] Kirsten Martin. 2019. Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160, 4 (2019), 835–850.

[246] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561* (2019).

[247] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561* (2019).

[248] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[249] John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine* 27, 4 (2006), 12–12.

[250] H Brendan McMahan et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning* 14, 1 (2021).

[251] Frank McSherry and Ilya Mironov. 2009. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 627–636.

[252] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 94–103.

[253] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[254] A. Menon and R. Williamson. 2017. The cost of fairness in classification. *ArXiv* abs/1705.09055 (2017).

[255] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 107–118.

[256] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650* (2019).

[257] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[258] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* 19, 6 (2018), 1236–1246.

[259] Sparsh Mittal and Jeffrey S Vetter. 2014. A survey of methods for analyzing and improving GPU energy efficiency. *ACM Computing Surveys (CSUR)* 47, 2 (2014), 1–23.

[260] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[261] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 19–38.

[262] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[263] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1765–1773.

[264] Laura Moy. 2019. How police technology aggravates racial inequity: A taxonomy of problems and a path forward. *Available at SSRN* 3340898 (2019).

[265] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1101–1111.

[266] Razieh Nabi and I. Shpitser. 2018. Fair Inference on Outcomes. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* 2018 (2018), 1931–1940.

[267] Valeria Nikolaenko, Stratis Ioannidis, Udi Weinsberg, Marc Joye, Nina Taft, and Dan Boneh. 2013. Privacy-preserving matrix factorization. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 801–812.

[268] Tong Niu and Mohit Bansal. 2018. Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models. *arXiv preprint arXiv:1809.02079* (2018).

[269] Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. 2021. An Empirical Study on the Relation Between Network Interpretability and Adversarial Robustness. *SN Computer Science* 2, 1 (2021), 1–13.

[270] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[271] Future of Life Institute. 2017. Asilomar AI principles. https://futureoflife.org/ai-principles/ Accessed March 18, 2021.

[272] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[273] Pete Pachal. 2015. Google Photos Identified Two Black People as' Gorillas'. *Mashable, July* 1 (2015).

[274] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 446–457.

[275] Nicolas Papernot, Ian Goodfellow, Ryan Sheatsley, Reuben Feinman, and Patrick McDaniel. 2016. cleverhans v1.0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768* (2016).

[276] Angshuman Parashar, Priyanka Raina, Yakun Sophia Shao, Yu-Hsin Chen, Victor A Ying, Anurag Mukkara, Rangharajan Venkatesan, Brucek Khailany, Stephen W Keckler, and Joel Emer. 2019. Timeloop: A systematic approach to dnn accelerator evaluation. In *2019 IEEE international symposium on performance analysis of systems and software (ISPASS)*. IEEE, 304–315.

[277] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).

[278] Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 742–751.

[279] Geoff Pleiss, M. Raghavan, Felix Wu, J. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *NIPS*.

[280] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923* (2020).

[281] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* (2019), 1–19.

[282] Kristina Preuer, Günter Klambauer, Friedrich Rippmann, Sepp Hochreiter, and Thomas Unterthiner. 2019. Interpretable deep learning in drug discovery. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 331–345.

[283] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779* (2019).

[284] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1, 1 (1986), 81–106.

[285] Evani Radiya-Dixit and Florian Tramer. 2021. Data Poisoning Won't Save You From Facial Recognition. In *ICML 2021 Workshop on Adversarial Machine Learning*.

[286] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344* (2018).

[287] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[288] Waseem Rawat and Zenghui Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* 29, 9 (2017), 2352–2449.

[289] 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.

[290] Leslie Rice, Eric Wong, and Zico Kolter. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*. PMLR, 8093–8104.

[291] 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 digital medicine* 3, 1 (2020), 1–7.

[292] Maria Rigaki and Sebastian Garcia. 2020. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646* (2020).

[293] Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. 2013. Learning separable filters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2754–2761.

[294] James A Rodger and Parag C Pendharkar. 2004. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies* 60, 5-6 (2004), 529–544.

[295] Crefeda Faviola Rodrigues, Graham Riley, and Mikel Luján. 2018. SyNERGY: An energy measurement and prediction framework for Convolutional Neural Networks on Jetson TX1. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*. The Steering Committee of The World Congress in Computer Science, Computer …, 375–382.

[296] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[297] Adam Rose. 2010. Are face-detection cameras racist. *Time Business* 1 (2010).

[298] Bita Darvish Rouhani, M Sadegh Riazi, and Farinaz Koushanfar. 2018. Deepsecure: Scalable provably-secure deep learning. In *Proceedings of the 55th Annual Design Automation Conference*. 1–6.

[299] Benjamin IP Rubinstein and Francesco Alda. 2017. diffpriv: An R Package for Easy Differential Privacy. (2017). https://github.com/brubinstein/diffpriv.

[300] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[301] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach. (2002).

[302] Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. 2017. Improving smiling detection with race and gender diversity. *arXiv preprint arXiv:1712.00193* 1, 2 (2017), 7.

[303] Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. 2020. Adversarial attacks on copyright detection systems. In *International Conference on Machine Learning*. PMLR, 8307–8315.

[304] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: what it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/ISPA*, Vol. 1. IEEE, 57–64.

[305] Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. 2009. Efficient privacy-preserving face recognition. In *International Conference on Information Security and Cryptology*. Springer, 229–244.

[306] Paul Sajda. 2006. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.* 8 (2006), 537–565.

[307] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[308] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.

[309] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[310] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[311] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.

[312] 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.

[313] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*. 6103–6113.

[314] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *arXiv preprint arXiv:1904.12843* (2019).

[315] Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *arXiv preprint arXiv:1912.11078* (2019).

[316] Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5248–5264.

[317] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 1589–1604.

[318] Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. 2016. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*. PMLR, 2217–2225.

[319] 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. *Scientific reports* 10, 1 (2020), 1–12.

[320] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3407–3412.

[321] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[322] K Shyong, Dan Frankowski, John Riedl, et al. 2006. Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In *International Conference on Emerging Trends in Information and Communication Security*. Springer, 14–29.

[323] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[324] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430* (2018).

[325] Nathalie A Smuha. 2019. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International* 20, 4 (2019), 97–106.

[326] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 241–257.

[327] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 245–248.

[328] Dimitrios Stamoulis, Ting-Wu Chin, Anand Krishnan Prakash, Haocheng Fang, Sribhuvan Sajja, Mitchell Bognar, and Diana Marculescu. 2018. Designing adaptive neural networks for energy-constrained image classification. In *Proceedings of the International Conference on Computer-Aided Design*. 1–8.

[329] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591* (2019).

[330] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[331] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355* (2019).

[332] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

[333] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[334] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136* (2019).

[335] Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, and Songcan Chen. 2021. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems* 34 (2021).

[336] R Tatman. 2016. Google's speech recognition has a gender bias. *Making Noise and Hearing Things* 12 (2016).

[337] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2020. Trustworthy artificial intelligence. *Electronic Markets* (2020), 1–18.

[338] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[339] 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 *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.

[340] Florian Tramer and Dan Boneh. 2021. Differentially Private Learning Needs Better Features (or Much More Data). In *International Conference on Learning Representations*. https://openreview.net/forum?id=YTWGvpFOQD-

[341] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 601–618.

[342] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* (2018).

[343] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400* (2014).

[344] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. 2019. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* 18, 6 (2019), 463–477.

[345] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088* (2019).

[346] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[347] Tao Wang, Zhigao Zheng, A Bashir, Alireza Jolfaei, and Yanyan Xu. 2020. Finprivacy: A privacy-preserving mechanism for fingerprint identification. *ACM Transactions on Internet Technology* (2020).

[348] Yuxin Wang, Qiang Wang, Shaohuai Shi, Xin He, Zhenheng Tang, Kaiyong Zhao, and Xiaowen Chu. 2020. Benchmarking the performance and energy efficiency of ai accelerators for ai training. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*. IEEE, 744–751.

[349] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8919–8928.

[350] Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 309 (1965), 63–69.

[351] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security* 15 (2020), 3454–3469.

[352] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.

[353] Maranke Wieringa. 2020. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 1–18.

[354] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal* 1, 2 (2017), 1–12.

[355] David H Wolpert and William G Macready. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1, 1 (1997), 67–82.

[356] Eric Wong and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*. PMLR, 5286–5295.

[357] Eric Wong, Leslie Rice, and J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. arXiv:2001.03994 [cs.LG]

[358] Eric Wong, Frank Schmidt, and Zico Kolter. 2019. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*. PMLR, 6808–6817.

[359] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems* 33 (2020).

[360] Kaiwen Wu, Allen Wang, and Yaoliang Yu. 2020. Stronger and faster Wasserstein adversarial attacks. In *International Conference on Machine Learning*. PMLR, 10377–10387.

[361] Yannan Nellie Wu, Joel S Emer, and Vivienne Sze. 2019. Accelergy: An architecture-level energy estimation methodology for accelerator designs. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 1–8.

[362] D. Xu, S. Yuan, L. Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. *2018 IEEE International Conference on Big Data (Big Data)* (2018), 570–575.

[363] Han Xu, Xiaorui Liu, Yaxin Li, and Jiliang Tang. 2020. To be Robust or to be Fair: Towards Fairness in Adversarial Training. *arXiv preprint arXiv:2010.06121* (2020).

[364] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. 2020. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing* 17, 2 (2020), 151–178.

[365] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. 2021. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research* 5, 1 (2021), 1–19.

[366] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-Adversarial Dialogue for Safe Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2950–2968.

[367] Mingfu Xue, Jian Wang, and Weiqiang Liu. 2021. DNN intellectual property protection: Taxonomy, attacks and evaluations. In *Proceedings of the 2021 on Great Lakes Symposium on VLSI*. 455–460.

[368] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[369] Tien-Ju Yang, Yu-Hsin Chen, and Vivienne Sze. 2017. Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5687–5695.

[370] Andrew C Yao. 1982. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 160–164.

[371] Catherine Yeo and Alyssa Chen. 2020. Defining and Evaluating Fair Natural Language Generation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. 107–109.

[372] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 32 (2019), 9240.

[373] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially Private Fine-tuning of Language Models. arXiv:2110.06500 [cs.LG]

[374] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. 2021. Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=7aogOj_VYO0

[375] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. Large Scale Private Learning via Low-rank Reparametrization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12208–12218. http://proceedings.mlr.press/v139/yu21f.html

[376] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. Indiscriminate Poisoning Attacks Are Shortcuts. arXiv:2111.00898 [cs.LG]

[377] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953* (2018).

[378] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 430–438.

[379] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in Graph Neural Networks: A Taxonomic Survey. *arXiv preprint arXiv:2012.15445* (2020).

[380] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[381] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[382] R. Zemel, Ledell Yu Wu, Kevin Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In *ICML*.

[383] Fadila Zerka, Samir Barakat, Sean Walsh, Marta Bogowicz, Ralph TH Leijenaar, Arthur Jochems, Benjamin Miraglio, David Townend, and Philippe Lambin. 2020. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *JCO clinical cancer informatics* 4 (2020), 184–200.

[384] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[385] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. *arXiv preprint arXiv:2004.14088* (2020).

[386] 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.

[387] L. Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI*.

[388] Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Lizhen Cui, and Xiangliang Zhang. 2021. Graph Embedding for Recommendation against Attribute Inference Attacks. *arXiv preprint arXiv:2101.12549* (2021).

[389] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.

[390] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.

[391] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).

[392] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 253–261.

[393] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.

[394] Zhe Zhang and Daniel B Neill. 2016. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292* (2016).

[395] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610* (2020).

[396] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

[397] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[398] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

[399] 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.

[400] Yingxue Zhou, Steven Wu, and Arindam Banerjee. 2021. Bypassing the Ambient Dimension: Private {SGD} with Gradient Subspace Identification. In *International Conference on Learning Representations*. https://openreview.net/forum?id=7dpmlkBuJFC

[401] Ligeng Zhu and Song Han. 2020. Deep leakage from gradients. In *Federated Learning*. Springer, 17–31.

[402] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

[403] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair.

[404] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2847–2856.

[405] Daniel Zügner and Stephan Günnemann. 2019. Adversarial attacks on graph neural networks via meta learning. *arXiv preprint arXiv:1902.08412* (2019).

[406] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial Attacks on Neural Networks for Graph Data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Jul 2018). https://doi.org/10.1145/3219819.3220078