Management System Engineering

REVIEW ARTICLE

Open Access

A brief review on algorithmic fairness

Xiaomeng Wang[1,2*], Yishi Zhang[1,2] and Ruilin Zhu[3]

Abstract

Machine learning algorithms are widely used in management systems in different fields, such as employee recruitment, loan provision, disease diagnosis, etc., and even in some risky decision-making areas, playing an increasingly crucial role in decisions affecting people's lives and social development. However, the use of algorithms for automated decision-making can cause unintentional biases that lead to discrimination against certain specific groups. In this context, it is crucial to develop machine learning algorithms that are not only accurate but also fair. There is an extensive discussion of algorithmic fairness in the existing literature. Many scholars have proposed and tested definitions of fairness and attempted to address the problem of unfairness or discrimination in algorithms. This review aims to outline different definitions of algorithmic fairness and to introduce the procedure for constructing fair algorithms to enhance fairness in machine learning. First, this review divides the definitions of algorithmic fairness into two categories, namely, awareness-based fairness and rationality-based fairness, and discusses existing representative algorithmic fairness concepts and notions based on the two categories. Then, metrics for unfairness/discrimination identification are summarized and different unfairness/discrimination removal approaches are discussed to facilitate a better understanding of how algorithmic fairness can be implemented in different scenarios. Challenges and future research directions in the field of algorithmic fairness are finally concluded.

Keywords: Algorithmic fairness, Fairness definition, Fairness identification, Unfairness removal, Causal inference

1 Introduction

Machine learning algorithms have been widely used and have become increasingly important in automated decision-making systems in business and government (Zhang 2018; Lambrecht and Tucker 2019; Teodorescu et al. 2021; Kallus et al. 2022). With the advantage of processing massive information and seemingly fair output, algorithms were believed to be successful in supporting decision-making. However, this is unfortunately not the case since machine learning algorithms are not always as objective as we would expect. Algorithms are vulnerable to biases that render their decisions "unfair" (Verma 2019). A biased model may inadvertently encode human prejudice due to biases in data (Mehrabi et al. 2021). Specifically, the algorithm may be discriminatory when it learns incorrect patterns, like stereotypes, from the observed data to make predictions and affect people's lives (Kallus et al. 2022). Furthermore, the algorithm itself may also lead to algorithm unfairness/discrimination (Danks and London 2017). The algorithm may sacrifice high performance on minority groups to achieve higher accuracy on overall samples while putting minority groups in a disadvantageous position. A typical case of algorithm discrimination is that COMPAS measures the risk of a person recommitting another crime and falsely links African-American offenders with high-risk recidivist scores (Chouldechova 2017). Besides, similar problems have been found in employment, insurance, and advertising. In another case of a hiring application, it was recently exposed that Amazon discovered that their automated hiring system based on machine learning was discriminating against female candidates, particularly for software development and technical positions. One

*Correspondence: wangxiaomeng@whut.edu.cn

[1] School of Management, Wuhan University of Technology, Wuhan 430070, China
Full list of author information is available at the end of the article

suspected reason for this is that most recorded historical data were for male software developers[1].

Fairness means dealing with things reasonably and not taking sides. Fair machine learning algorithms refer to no bias or preference for individuals or groups due to their inherent or acquired attributes in the decision-making process (Saxena et al. 2019). Since many automated decisions (including which individuals will receive jobs, loans, medication, bail, or parole) can significantly impact people's lives, there is great importance in assessing and improving the ethics of the decisions made by these automated systems (Carey and Wu 2022). The fairness of the outputs is not only the evaluation of the algorithm performance but also affects the benefit distribution in the real decision-making situation. Thus, building a reasonable model to ensure fair decision-making of algorithms is of great theoretical significance and application value. ACM (the American Computer Society) started to set up a FAccT conference that discussed the issues of fairness, accountability, and transparency in cross-domain fields including computer science, statistics, law, social science, and humanities in 2018. In addition, several important international conferences on artificial intelligence, including ICML, NeurIPS, and AAAI, specially set up research topics to discuss fair machine learning (Niu et al. 2021; Yang et al. 2020).

This review aims to sort out the current state of the art of fairness in machine learning and to provide reference ideas for follow-up research. The key questions of fair machine learning research are how to establish a fair definition guided by law, ethics, and sociology, and how to design a fair machine learning algorithm driven by the fairness definition (Teodorescu et al. 2021; Carey and Wu 2022). Although various fairness definitions have been proposed, they are incompatible and cannot be used together. This article outlines different definitions of algorithmic fairness and provides a framework for constructing fair algorithms.

The main contributions of this article are as follows: we categorize definitions of fairness in the existing literature into two streams: awareness-based fairness and rationality-based fairness, where the latter contains most of the prevailing fairness notions that are categorized in the existing literature as "statistical-based awareness" and "causality-based awareness". We suggest viewing different definitions of fairness from both rationality and awareness perspectives, to avoid the conflict of different fairness metrics, inspiring researchers to explore fairness issues in both technical application and ethical aspects. We also summarize the process of the algorithmic fairness task into four stages: initialization, fairness definition, fairness identification, and unfairness/discrimination removal, which provides a feasible reference for constructing fair models in various application domains. Finally, we emphasize causal fairness definitions and present emerging trends in most recent research to guide subsequent researchers to research and explore algorithmic fairness.

The rest of this paper is structured as follows. Section 2 presents a roadmap of fairness in machine learning algorithms and introduces the processing flow of the fairness task of the algorithm from the overall perspective. Section 3 introduces stage 1 in the roadmap. Section 4 discusses the criteria of fairness and its feasibility in practical implementation. Section 5 describes unfairness or discrimination detection approaches. Section 6 reviews possible solutions to remove unfairness in different scenarios. Several mechanisms are compared and their strengths and weaknesses are emphasized. Section 7 provides concluding remarks and sketches several open challenges for future research.

2 Roadmap for algorithmic fairness

Automated methods of algorithmic fairness analysis come from the field of bias analysis. The relationship between bias analysis and fairness analysis is analogous to that of physics to engineering. That is, bias analysis, at its core, emphasizes advancing statistical theory and often focuses on the fitness or the accuracy of estimation as an end in itself (Cheng et al. 2021). Computer-assisted or automated fairness analysis, on the other hand, refers to a set of techniques that use computing or statistical power to answer questions of fairness in market and business (Lambrecht and Tucker 2019; Zhang 2018; Zhang et al. 2019; Kallus et al. 2022), politics and law (Teodorescu et al. 2021; Chen et al. 2021), and public affairs (Editorial. 2016; Barocas and Selbst 2016; Caton and Haas 2020). In these fields, fairness represents some focal structure of interests, and computers are used to measure fairness, provide efficient and systematic comparisons, and sometimes detect unfairness/discrimination that neither practitioners nor researchers can be easily aware of. In other words, while bias analysis is a research topic that is primarily concerned with bias in the data, for managerial researchers and social practitioners, fairness analysis is merely a lens through which to view human's thought, behavior, and even the conflict of interest. Analyzing fairness, in many contexts, is not the ultimate goal of the practitioners and the researchers, but is instead a precursor for making socially responsible decisions where the stakeholders are involved.

[1] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secretai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Therefore, we use the term "fairness analysis" over "bias analysis" and "algorithmic unfairness/discrimination" over "algorithmic bias" in this paper. Although we follow convention by using the term "automated", this should not imply that human intervention is absent. In fact, many of the tasks-particularly the definition of fairness-are iterative processes that require human design, modification, and interpretation. In the following sections, we discuss the design and execution of automated fairness analysis in detail, beginning with selection of initialization and connecting statistical and causal aspects to people's perceptions about fairness in different contexts of this society.

Without ambiguity, we use "vulnerable group (non-vulnerable group)", "protected group (unprotected group)", and "unprivileged group (privileged group)" interchangeably, and "sensitive attribute" and "protected attribute" interchangeably in this review.

3 Initialization

It is ubiquitous in the real world that the prediction/decision outcomes are sensitive to the stakeholders from different groups. Under these circumstances, an unignorable task is to judge whether the outcomes are fair, and is especially true when the prediction/decision is made by automated methods like machine learning algorithms (Zhang et al. 2016b). Intuitively, the stakeholders in a specific group (e.g., people with certain gender or race) would probably make comparisons to ones in other group(s), in such a way as to be aware of whether they are treated the same. The result of such comparisons belong to a perceptual cognition of fairness. It commonly appears in informal scenarios or impromptu situations involving the distribution of benefits, or the cases where individual feelings play an important role. We call this kind of fairness "awareness-based fairness", which mainly involves fairness through unawareness (i.e., totally excluding sensitive variables like gender or race that affect fairness judgments) and fairness through awareness (Kusner et al. 2018; Zhang 2018).

In contrast, some techniques for rational analysis, e.g., statistical tools or causal analytical ones, would be applied in fairness analysis to pursue more scientific and reasonable judgments and the subsequent solutions. Fairness notions defined using such techniques are mostly group-oriented, and the conclusions drawn tend to have a global meaning and are often more appropriate for management of society, market, and law. However, the deficiency of them is the ignorance of individual perceptions which makes them sometimes conflict with individual perceptions of fairness (Teodorescu et al. 2021). We call this type of fairness "rationality-based fairness", which mainly includes two main camps: statistical-based fairness and causality-based fairness (Kusner et al. 2018; Carey and Wu 2022).

The choice of definition of fairness depends entirely on the specific situation at hand (different positions and roles, individual- or group-oriented, formal or informal, etc.). In a legal situation, for example, if you feel you have been treated unfairly, the content of your claim may be that someone similar to you has been treated quite differently. But for a judge to decide whether a decision maker has made a discriminatory decision, he/she often needs to conduct a thorough and careful investigation from the perspective of the group. To some extent, the relationship between awareness-based fairness and rationality-based fairness is like that of scientific decision-making and decision-making behavior which is characterized by finite rationality or even irrationality. As we write this review, researchers are now developing new fairness notions in an attempt to reconcile the conflicts/contradictions among multiple aspects (rational and emotional, group and individual, etc.), where we believe the fairness notion via causality is the one that has the most potential to come close to this goal.

In the following sections, we will go through the rest stages shown in Fig. 1, to review the representative work from the perspective of the whole process of algorithmic fairness, including fairness definition, fairness identification, unfairness/discrimination removal (if necessary), and finally obtaining fair prediction/decision outcomes.

4 Fairness definition

In the real world, different machine learning tasks focus on different issues, so it is difficult to determine a general definition of fairness. This section summarizes the definitions of fairness proposed in the existing literature. For awareness-based fairness, according to whether sensitive attributes are considered, it can be categorized as fairness through awareness and fairness through unawareness. For rationality-based fairness, according to the role of protected attributes as well as the mathematical paradigm applied in the process of building fair machine learning algorithms, it can be roughly divided into two categories: statistical-based fairness and causality-based fairness. Table 1 concludes all kinds of fairness measurements discussed in this paper. Detailed definitions will be illustrated in the following subsections.

4.1 Awareness-based fairness

4.1.1 Fairness through unawareness

Fairness through unawareness is an intuitive definition of fairness. It is a perception- (rather than rationality-) oriented definition. Fairness through unawareness focuses on how to directly deal with sensitive attributes to obtain fairness. If sensitive attributes are not explicitly used in

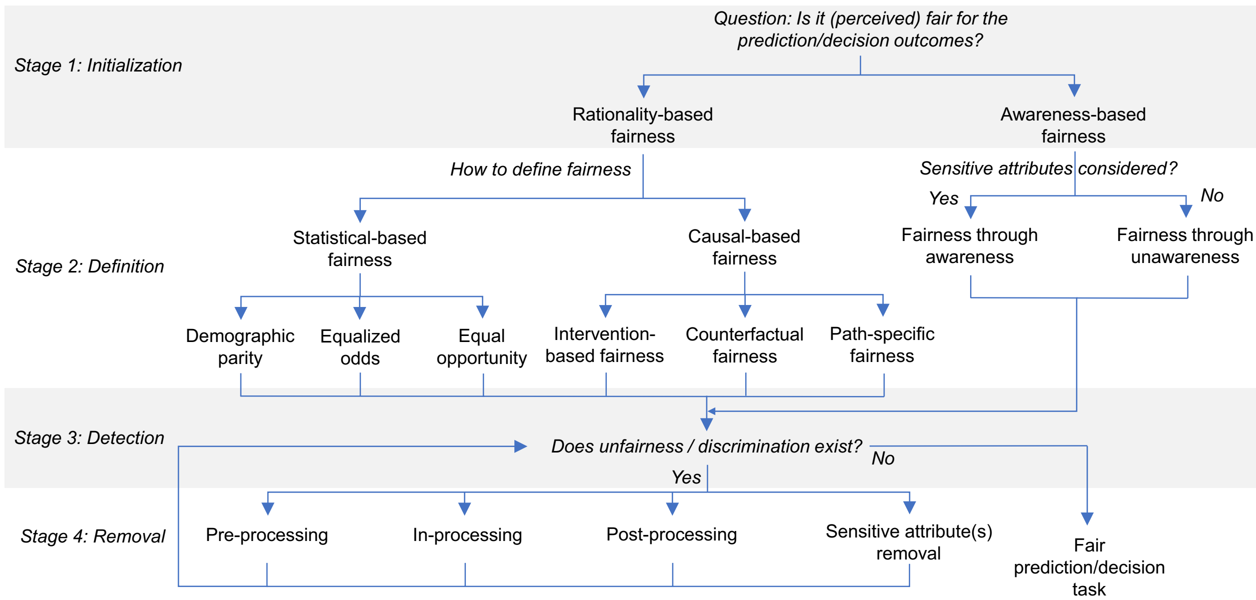

Fig. 1 Stages of automated algorithmic fairness analysis

Table 1 A summary of typical algorithmic fairness notions

Category	Notion		Definition	References
Awareness-based fairness	Fairness through unawareness		$\hat{y} = \mathcal{F}(\mathbf{x}), S \notin \mathbf{X}$	Chen et al. (2019); Brown et al. (2016); Zhang (2018); Kallus et al. (2022)
	Fairness through awareness		$d_1((\mathbf{x}_i, s_i), (\mathbf{x}_j, s_j)) \leq d_2(\hat{y}_i, \hat{y}_j)$	Zhang et al. (2016b)
Rationality-based fairness	Statistical-based Fairness	Demographic parity	$P(\hat{y}\|s = 0) = P(\hat{y}\|s = 1)$	Dwork et al. (2012)
		Equalized odds	$P(\hat{y} = 1\|s = 0, y = 0) = P(\hat{y} = 1\|s = 1, y = 0)$	Hardt et al. (2016)
		Equality of opportunity	$P(\hat{y} = 1\|s = 0, y = 1) = P(\hat{y} = 1\|s = 1, y = 1)$	Hardt et al. (2016)
		Test Fairness	$P(y = 1\|s = 0, \hat{y}) = P(y = 1\|s = 1, \hat{y})$	Chouldechova (2017)
	Causal-based Fairness	Intervention-based fairness	$P(\hat{y}\|do(s = 0)) = P(\hat{y}\|do(s = 1))$	Loftus et al. (2018); Khademi et al. (2019)
		Path-specific fairness	$P(\hat{y}_{s=0}\|\boldsymbol{\pi}) = P(\hat{y}_{s=1}\|\boldsymbol{\pi})$	Wu et al. (2019b); Chiappa (2019); Zhang et al. (2019)
		Counterfactual fairness	$P(\hat{y}_{s=0}\|s = 0, \mathbf{x}) = P(\hat{y}_{s=1}\|s = 0, \mathbf{x})$	Kusner et al. (2018); Wu et al. (2019a)

the decision-making process, then the algorithm achieves fairness through unawareness (Zhang 2018; Chen et al. 2019; Kallus et al. 2022). Let $\mathcal{F}(\cdot)$ be the learning process of the algorithm, $\mathbf{X}$ be the attributes in the dataset, S be the sensitive attribute, and $\hat{Y}(Y)$ be the predicted outcome (the ground truth) $S \notin \mathbf{X}$ ($\mathbf{x}, s$, and $\hat{y}$ are the value assignments of $\mathbf{X}$, S,and $\hat{Y}$, respectively), then

$$\hat{y} = \mathcal{F}(\mathbf{x}), S \notin \mathbf{X} \tag{1}$$

implies that the learning process neglects the sensitive attribute and the outcome $\hat{y}$ is perceived fairness. Although fairness through unawareness is simple and intuitive, it may introduce indirect unfairness when other attributes are highly related to the protected attribute (e.g., street and zipcode). If these attributes are used by the algorithms, the outcomes may still be unfair while giving the impression that the algorithms act fairly

(Teodorescu et al. 2021). It would only be applicable in the unlikely scenario of no correlation between the sensitive attribute and the rest attributes used to predict outcomes (Teodorescu et al. 2021).

4.1.2 Fairness through awareness

Fairness through awareness defines fairness via the viewpoint of individuals (Zhang et al. 2016b). If individuals with similar value assignments of the attributes including the sensitive attribute (which means they are similar to each other, e.g., with similar preferences, characteristics, experiences, etc.) are treated similarly, then the algorithm achieves fairness through awareness/individual fairness (Dwork et al. 2012; Luong et al. 2011). To effectively measure the similarity of the attributes as well as the outcome, two corresponding similarity/distance functions should be elaborately defined to make this fairness notion practical. Let $\mathbf{Z}$ be the attributes in the dataset, $\mathbf{X} \subset \mathbf{Z}$ is a subset of attributes excluding S, fairness through awareness can be formally expressed as

$$d_1((\mathbf{x}_i, s_i), (\mathbf{x}_j, s_j)) \leq d_2(\hat{y}_i, \hat{y}_j) \tag{2}$$

where $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ denote the distance functions, and subscripts i and j denote two individuals (samples in the dataset). Eq.(2) implies the perceived differentiation of the outcomes of two individuals should not be greater than the discrepancy in their attributes.

Although this definition sounds reasonable, it is difficult to realize because it is challenging to measure the distance between individuals under specific tasks. Because it is almost impossible to obtain enough fine-grained features of individuals in real situations, i, j are more likely to appear in the form of groups in the data. Thus, Eq. (2) cannot guarantee the protected and unprotected groups are being treated fairly, which requires more rational notions of fairness to be proposed.

4.2 Rationality-based fairness

4.2.1 Statistical-based fairness

Statistical-based fairness requires that the protected group be treated similarly to the non-vulnerable group or the whole group (Lum and Johndrow 2016). Taking the famous algorithm COMPAS as an example, the race is regarded as a protected attribute, and the algorithm's performance across different race groups can be a sign to determine when the output is fair. ProPublica[2] reveals the differences in the false positive rate and false negative rate of the risk assessment results between the European-American defendant group and the African-American defendant group. Specifically, the European-American defendant group is less (more) likely to be marked as high (low) risk even when they actually have the same probability of recommitting crimes. This violates statistical-based fairness. Statistical fairness does not need to make additional assumptions on the data (Pessach and Shmueli 2023) and is easy to verify, but this definition cannot guarantee fairness at the individual level (Makhlouf et al. 2022). According to the different contexts of usage, the existing statistical-based fairness can be divided into demographic parity and statistical fairness given the ground truth, and the latter can further be divided into equalized odds (Hardt et al. 2016; Mehrabi et al. 2021), equality of opportunity (Hardt et al. 2016; Zafar et al. 2017a), and test fairness (Kleinberg et al. 2016; Chouldechova 2017; Caton and Haas 2020).

Demographic parity Demographic parity (Corbett-Davies et al. 2017; Feldman et al. 2015; Kamishima et al. 2012), also known as statistical parity, requires protected and unprotected groups to obtain the same output prediction results with the same probability. If the output Y is independent of the protected attribute S in any case, then Y satisfies statistical parity, namely

$$P(\hat{y}|s=0) = P(\hat{y}|s=1) \tag{3}$$

This definition requires different groups to obtain the same output results with the same probability.

However, the effectiveness of the above definition will be weakened when S and Y are highly related. The distributions of other attributes are different between the protected group and other groups, and the final decision-making results associated with this definition may violate common sense in reality. To this end, the statistical fairness given the ground truth Y is introduced below, which additionally considers data marking on the basis of demographic fairness.

Statistical fairness given the ground truth Statistical fairness based on the ground truth (Caton and Haas 2020) measures the difference of error rate and correct rate of output results of each group and requires the difference to be minimized. As mentioned previously, it generally consists of equalized odds, equality of opportunity, and test fairness.

Equalized odds (Hardt et al. 2016) looks at the independence of the score and the sensitive variable conditional on the value of the target variable Y (i.e., the outcome). It computes the difference between the false-positive rates (FPRs), and the difference between the true-positive rates (TPRs) of the two groups. Equalized odds enforces equality of error rates across the sensitive attribute and the outcome, providing a stronger group fairness metric than demographic parity. Equalized odds has the following form

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

$$P(\hat{y}|s=0,y)=P(\hat{y}|s=1,y) \qquad \hat{y}=0,1,y=0,1 \tag{4}$$

The above definition of Equalized odds implies that each sensitive attribute requires an additional test of the criterion. This would be challenging for cases containing more than one sensitive attribute.

Equality of opportunity (Hardt et al. 2016; Zafar et al. 2017a) is similar with Equalized Odds but focuses on TPRs only. It is a weaker version of equalized odds, which can be described as

$$P(\hat{y}=1|s=0,y=1)=P(\hat{y}=1|s=1,y=1) \tag{5}$$

Similar to equalized odds and equality of opportunity, treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories.

Test fairness Test fairness (Chouldechova 2017) is a representative definition of calibration statistical fairness (Kleinberg et al. 2016; Chouldechova 2017). It states that for any predicted probability score $\hat{y}$, people in both protected and unprotected groups must have an equal probability of correctly belonging to the positive class:

$$P(y=1|s=0,\hat{y})=P(y=1|s=1,\hat{y}) \tag{6}$$

Notably, following the equality in terms of only one type of error (e.g., true positives) will increase the disparity in terms of the other error (Pleiss et al. 2017). Arguments about statistical fairness recognize that these criteria are based purely on probabilistic independence. Potential spurious relations between sensitive attributes and outcomes may lead to misunderstanding of unfairness (Kusner et al. 2018).

4.2.2 Causality-based fairness

Statistical-based fairness notions are correlation-based (Wu et al. 2018; Zhang et al. 2017, 2019) and attempt to pursue “literal equity” in the outcome only according to the protected attribute, e.g., demographical parity requires that the proportion of positive outcome (e.g. admission) is the same for all sub-populations (e.g. male and female groups), and equal opportunity requires that the true positive rate (TPR) is the same for all sub-populations. They ignore the fact that “equity” is actually a result of the equilibrium of interest relations (Beretta et al. 2019; Gelfand et al. 2002). Causality-based fairness is different from statistical-based one in that it defines the causal effect of sensitive attribute on outcome as unfairness/discrimination (Carey and Wu 2022). Besides, it is not completely driven by the observational data, but requires additional causal relationships that reflect the principles of the socio-economic system and the knowledge of behaviors of the stakeholders.

To the best of our knowledge, most of the causality-based fairness notions are defined in the context of structural causal model [(SCM, (Judea 2009)], aiming to discover and eliminate the causal effect of sensitive attributes on outputs by intervening SCM (Kilbertus et al. 2020; Kusner et al. 2018; Zhang et al. 2016a; Nabi et al. 2018). SCM includes causal structure equations and a corresponding causal graph (Pearl 2009). Causal graph is a directed acyclic graph that represents causality among attributes. Nodes in a causal graph represent attributes, arrows indicate causality, and attribute nodes representing causes point to attribute nodes representing effects. *Do*-calculus is a technique of SCM to obtain the causal diagram after the intervention via only the observational data (Pearl et al. 2016). For example, the intervention on protected attribute S means to delete all the arrows pointing to S in the causal graph and assign a specific value to S, thus obtaining the causal graph after the intervention. Usually, $do(s=0)$ is used to indicate intervention on S, and attribute S is assigned a value of 0. This is very useful for fairness analysis using only observational data (i.e., like the paradigm of the statistical-based fairness analysis) in which sensitive attributes like gender and race are difficult to manipulate.

Causality-based fairness focuses on the causal relationship between sensitive attributes and results, and can specifically eliminate unfair effects in the system, while retaining fair parts. Causality-based fairness will use symbols like $y_{s=0}$ to represent the counterfactual predicted label (i.e., the counterfactual outcome) if s had been assigned a specific value 0 (which implies $s=1$ in the real-world)[3]. This notation is equivalent to $y|do(s=0)$ if S is still undetermined (i.e., $P(y_{s=0})=P(y|do(s=0))$. Fairness notions proposed from a causal perspective include intervention-based fairness (Loftus et al. 2018; Huang et al. 2020), path-specific fairness (Wu et al. 2019b; Chiappa 2019), and counterfactual fairness (Kusner et al. 2018; Garg et al. 2019; Wu et al. 2019a; Niu et al. 2021).

Intervention-based fairness Intervention-based fairness is the most natural definition of causality-based fairness (Loftus et al. 2018; Huang et al. 2020; Khademi et al. 2019). It is also referred to as fairness based on total causal effect (Huan et al. 2020). The only difference between intervention fairness and statistical parity is that it relies on intervening rather than the given sensitive attributes values. It requires that the output result Y satisfy (Loftus et al. 2018):

[3] The concept is somewhat non-straightforward to understand and readers may refer to Pearl et al. (2016) for more details.

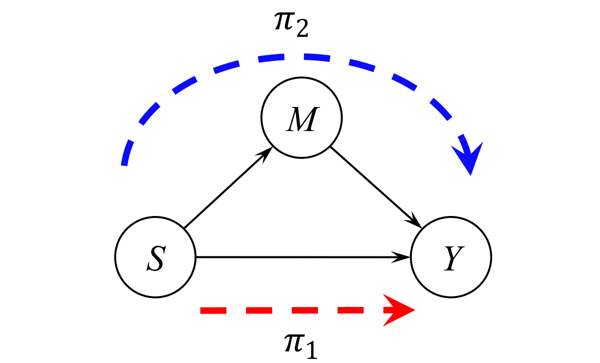

Fig. 2 A toy example of path-specific effects

$$P(\hat{y}|do(s=0)) = P(\hat{y}|do(s=1)) \quad (7)$$

Note that it is quite different from Eq. (3) though it can be estimated only on the observational data via some techniques (e.g., the frontdoor criterion from Judea (2009)) from SCM or the matching methods under the potential outcomes framework (Khademi et al. 2019; Huang et al. 2020). In addition, effectively using such techniques to implement Eq. (7) will require additional knowledge or assumptions of the causal structure of the attributes, in that some key information from the random controlled trial which is a direct way for do-calculus to conduct intervention is missing in observational data. Unfortunately, some specific structures containing confounders or mediators make it difficult to implement *do*-calculus from the observational data.

Path-specific fairness

To tackle the barrier mentioned above, fairness notions based on path-specific effect of SCM are proposed recently (Wu et al. 2019b; Chiappa 2019). It involves the knowledge of the causal structure of the attributes and labels and test whether it is unfair, i.e., measure the difference in the distribution of the prediction results of the same group after intervention, from the perspective of specific paths of the causal structure (Zhang et al. 2017).

$$P(\hat{y}_{s=0}|\boldsymbol{\pi}) = P(\hat{y}_{s=1}|\boldsymbol{\pi}) \quad (8)$$

where $\hat{y}_{s=0}$ is the counterfactual notion and $\boldsymbol{\pi}$ denotes the specific path in the causal structure. Compared with "literal fairness" observed from the data by statistical-based fairness notions, path-specific fairness is superior in some cases because it tries to distinguish different effects associated with various situations. For example, it can effectively identify the cause of gender discrimination from the example of graduate admissions at Berkeley (Bickel et al. 1975), where gender discrimination disappears when department choice that mediates the influences of gender on the admissions decision is considered.

In fact, the causal effects of S on Y through $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ shown in Fig.2 include direct and indirect effects, some of which indicate the unfairness/discrimination but some may not [(e.g., the explainable effect (Zhang et al. 2019)]. And whether it is associated with fairness depends on the interpretations that may represent different positions of the stakeholders.

However, since the counterfactual notion is intractable in most cases, even the direct effect from path-specific fairness, e.g., the natural direct effect (Pearl 2012b; Pearl and Mackenzie 2018) defined as

$$P(\hat{y}_{M=m}|do(s=0)) = P(\hat{y}_{M=m}|do(s=1)) \quad (9)$$

where M denotes the mediator between S and Y, is identifiable under some strong assumptions (Avin et al. 2005; Pearl 2012; Pearl et al. 2016). This restricts the applications of path-specific fairness notion in fairness analysis.

Counterfactual fairness Kusner et al. (2018) recognized that fairness should be regulated by explicitly modeling the causal structure of the world and thus proposed counterfactual fairness. The counterfactual fairness notion (Wu et al. 2019a; Garg et al. 2019; Niu et al. 2021) is based on the intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group. It can be described as follows:

$$P(\hat{y}_{s=0}|s=0,\mathbf{x}) = P(\hat{y}_{s=1}|s=0,\mathbf{x}) \quad (10)$$

where $\mathbf{x}$ is value assignments of $\mathbf{X}$ and we have $\mathbf{X} \subset \mathbf{Z}$ ($\mathbf{Z}$ is the attribute set excluding S).

The strength of this notion lies in that it can satisfy not only the awareness-based fairness analysis for individuals (it focuses on the counterfactual case for individuals) but also the rationality-based fairness analysis (it is defined in the context of SCM). However, it also faces the difficulty of intractability since it is conditioned on $s=1$ and at the same time depends on the counterfactual $s=0$, which is contradictory in practice.

The above Causality-based fairness notions rely mostly on SCM that may not be unique would encounter the problem of unidentifiability (Avin et al. 2005; Galles and Pearl 2013). To this end, there is a line of research (Zhang et al. 2019; Shpitser and Pearl 2007, 2012) trying to find the approximations instead of exact probabilities to make these notions applicable in practice.

Table 2 Unfairness/Discrimination removal methods

Method	Strength	Challenge
Pre-processing	Flexible to adapt to downstream tasks; Do not need to access the protected property when testing	The accuracy of that result needs to be guaranteed.
In-processing	It can balance the trade-off between algorithm accuracy and fairness; No access to protected properties	Dependent machine learning algorithm
Post-processing	Adapt to all kinds of algorithms	Need to access the protected attributes when testing the machine learning algorithm.

5 Fairness identification

In the task of fairness identification, the output of the algorithm needs to be judged according to the fairness notions discussed in stage 2.

5.1 Awareness-based fairness

For fairness through unawareness, neglecting sensitive attributes may not tackle the unfair problems because the rest attributes may have residual information about sensitive attributes, which may even exacerbate unfairness while giving the impression that the algorithms act fairly (Teodorescu et al. 2021). To address this challenge, some methods heuristically use proxy-based approaches, and optimization-based methods to predict and impute neglected sensitive attribute labels (Elliott et al. 2009; Hasnain-Wynia et al. 2012; Brown et al. 2016; Zhang 2018), although the validity of such methods still remains controversial (Kallus et al. 2022; Chen et al. 2019).

An identification approach that is widely used in fairness through awareness is to calculate the distances of samples or the distributions in the input and output spaces, requiring that similar individuals (similar input distances) receive the same treatment (similar output distances). Classifiers like k-nearest neighbor (kNN) can be applied to find the similar tuples (Luong et al. 2011). Recall that the notions of awareness-based fairness:

$$d_1((\mathbf{x}_i, s_i), (\mathbf{x}_j, s_j)) \le d_2(\hat{y}_i, \hat{y}_j)$$

where $d_1(\cdot, \cdot)$ and $d_2(\cdot, \cdot)$ denote the distance functions. To define the distance function, a distance metric is established to measure the per-attribute distance and the joint effect is obtained by summing up all the per-attribute distances. The normalized Manhattan distance and overlap measurement are widely used as the distance metrics (Luong et al. 2011; Zhang et al. 2016a).

5.2 Rationality-based fairness

As for rationality-based fairness, pursuing the probability distributions of the two groups to be exactly equal is unrealistic. A tractable method in practice is to calculate the difference or ratio of the outcome probabilities of the subgroups and consider that there is no (significant) unfairness/discrimination if it is less than a certain threshold (Caton and Haas 2020). Denote $\Theta(s)$ as the probability notations given the sensitive attribute $S = s$, the threshold-based identification approach for rationality-based fairness can be shown as

$$|\Theta(s = 0) - \Theta(s = 1)| \le \epsilon \tag{11}$$

or the fraction style

$$\frac{\Theta(s = 0)}{\Theta(s = 1)} \ge 1 - \epsilon \tag{12}$$

where ϵ denotes the fairness threshold. Formula (11) and (12) show an operational way for determine unfairness in practice, where $\epsilon = 0.8$ corresponds to the “four-fifths principle” in law (Adel et al. 2019; Zafar et al. 2017b) and thus is often selected in the identification process (Wu et al. 2019a).

6 Unfairness/discrimination removal

After any unfairness/discrimination has been detected in stage 3, it comes to the removal process to make the algorithms (or their outputs) discrimination-free. Mechanisms used to remove unfairness/discrimination is essentially interfering with algorithms, which can be categorized into pre-processing, in-processing, and post-processing ones. Table 2 compares different mechanisms for eliminating unfairness. Pre-processing mechanism aims to obtain unbiased datasets. In-processing mechanism achieves fairness by modifying the algorithms. Post-processing mechanism adjusts the outputs of the algorithms to make the decision fair. All these mechanisms will be further discussed in the following sections.

6.1 Pre-processing

Pre-processing approaches (Calmon et al. 2017; Feldman et al. 2015; Kamiran and Calders 2009, 2012) focus on dataset pre-processing, trying to adjust the datasets to eliminate the biases introduced by attributes S.

The unbiased datasets or processed original datasets conduce to improve the fairness of algorithms' outputs without modifying the machine learning algorithm. Pre-processing approaches can easily adapt to various downstream tasks, but may sacrifice accuracy and interpretability.

Feldman et al. (2015) modified all the non-protected attributes to ensure that protected attribute S cannot be predicted from the non-protected attributes. As a result, decision Y is determined by the non-protected attributes. Žliobaite et al. (2011) proposed the use of log-linear modeling to capture and measure discrimination and developed a method for discrimination prevention by modifying significant coefficients of the fitted log-linear model and generating unbiased datasets. Xu et al. (2020) proposed conditional fairness, which means outcome variables should be independent of sensitive attributes conditional on these fair variables. They proposed a Derivable Conditional Fairness Regularizer (DCFR), which can be integrated into any decision-making model, to track the trade-off between precision and fairness of algorithmic decision-making. Kamiran and Calders (2009) proposed a method based on massaging the dataset by making the least intrusive modifications which was used to build a Classification with No Discrimination (CND). Specifically, they used a ranking function learned on the biased data and modified training data based on this function.

6.2 In-processing

In-processing approaches (Bellamy et al. 2018; Calders and Verwer 2010; d'Alessandro et al. 2017; Kamishima et al. 2012) aims to change the training process of the algorithm (i.e., adding some fairnees constraints). Usually, one or more fairness metrics are incorporated into the model optimization functions to maximize both accuracy and fairness, providing a good view for the trade-off between fairness and accuracy. However, in-processing mechanism depends on specific algorithms. That is, different adjustment methods need to be proposed for different algorithms. For example, Kamiran et al. (2010) developed a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free. Zafar et al. (2017a) added the measure of fairness into the classification learning formulation as the constraint so that the classifier learned satisfies the fairness requirement. Chen et al. (2022) proposed an in-processing model for discrimination mitigation in natural language processing. Garg et al. (2019) proposed a model training scheme that can employ fairness constraints, which engaged fairness in cyberbullying detection algorithm.

6.3 Post-processing

Post-processing mechanism (Danks and London 2017; Hardt et al. 2016; Kamiran et al. 2010) concerns the fairness of decision results and tries to modify the algorithms' outputs. The advantage of the post-processing mechanism is that it does not interfere with the training process of the algorithms, which makes it applicable to different algorithms. However, modifying the outputs may reduce the accuracy of the algorithms, and it is still necessary to test whether the modified results are fair. Hardt et al. (2016) simply flipped outcomes of some samples so that the decision can meet equalized odds. But it will sacrifice the performance of algorithms. To solve this problem, Corbett-Davies et al. (2017) and Jung et al. (2017) imputed algorithm bias to its different performance on minority groups and majority groups. They similarly suggest selecting separate thresholds for each group separately, in a manner that maximizes accuracy and minimizes demographic parity. Dwork et al. (2012) proposed a decoupling technique to learn a different classifier for each group. They additionally combine a transfer learning technique with their procedure to learn from out-of-group samples.

A distinct advantage of pre- and post-processing approaches is that they do not modify the machine learning method explicitly. This means that (open source) machine learning libraries can be leveraged unchanged for model training. However, they do not directly control the optimization function of the machine learning model itself. Yet, modifying original data and/or model output may have legal implications (Barocas and Selbst 2016), and models still lack interpretability (Lepri et al. 2018; Lum and Johndrow 2016), which may be at odds with current data protection legislation with respect to interpretability. Only in-processing approaches can optimize notions of fairness during model training. However, this requires the optimization function to be either accessible, replaceable, and/or modifiable, which may not always be the case.

7 Conclusion

7.1 Summary

Algorithmic fairness has significance at the legal and social levels and is more of an interdisciplinary subject of social science and computer science. Fairness is a relative social concept and there is no fairness in an absolute sense. Fair machine learning algorithms gradually improve the fairness of machine learning algorithms by exploring the mechanisms to eliminate unfairness or discrimination. In this review, we have outlined different definitions of algorithmic fairness and provided a framework for constructing fair algorithms. We suggest viewing different definitions of fairness from both rationality and awareness perspectives, to avoid the conflict of different fairness metrics. We summarize the process of the algorithmic fairness task into four stages: initialization, fairness definition, fairness identification, and unfairness/discrimination removal. In future work, there is a need to deploy advanced fairness machine learning algorithms in various application domains and to develop unified and complete fairness metrics. Therefore, exploring fairness issues in both technical application and ethical aspects is necessary.

7.2 Future directions

7.2.1 Exploring the causal structure of data to strengthen fairness definitions

The causes of unfairness in machine learning algorithms are various and complex, and different biases have different influences on realistic applications. The very first challenge in fair machine learning is to provide a comprehensive definition of fairness. Whether an algorithm is fair not only depends on the model and data but also the task requirements.

As we mentioned above, various fairness notions are proposed in existing research. In addition, there is a lack of comprehensive and multi-dimensional algorithm fairness evaluation metrics and assessment systems to effectively quantify the fairness risk faced by machine learning algorithms, which makes it impossible to guarantee the fairness of machine learning models employed in different decision-making scenarios.

On the other hand, ignoring the causal structure in data may lead to the misuse of the definition of fairness. In the well-known Berkley example, the admission result of this college is considered unfair to females because the overall admission rate of males is higher than that of females. However, the situation is reversed when we compare the admission rates of different genders from the perspective of departments. The admission rate of women in almost every department is higher. In this example, the admission results are falsely related to gender due to personal choices, which leads to superficial discrimination. In many similar situations, the pseudo correlation between sensitive variables and results will affect the detection of discrimination. Thus, there is a causal structure that must be taken into account when detecting discrimination.

We deem that it is an important research trend to explore the causal structure of data exploiting causal inference techniques in the field of algorithm fairness. Introducing causal inference methods into algorithmic fairness can assist in building more convincible fairness notions. In the unfairness detection stage, it is crucial to understand the root causes of the problem when tackling the discrimination problem. In other words, it is necessary to determine whether sensitive attributes have an impact on the outcome, and how to eliminate the such impact. Causal inference can play a part in analyzing which types of discrimination should be allowed and which should not. Causal fairness notions and discrimination detection approaches, such as PSE and intervention-based fairness, are proposed to help solve these problems and more effort is needed in causal fair learning to improve fairness.

7.2.2 Bridging the gap between fairness notions and real applications

The application scenarios of machine learning models are multiple, and there may be difficulties in data collection in practical applications, which bring challenges to fair machine learning.

There are several barriers when applying fair machine learning in real scenarios. The significance of machine learning algorithms lies not only in fitting the distribution of the training set but also in fitting the distribution of the real world. Sensitive attributes are often inaccessible and difficult to test in real applications. The situation gets even worse when the training dataset is selection biased, which means it does not contain samples appearing in real world. Existing work uses proxies to solve the problem of inaccessible sensitive attributes. However, whether the proxy is fair enough is still worth talking in terms of training fair and accurate models.

Another obstacle is that existing fairness definitions may be inefficient and cannot adapt to the complex reality of human-machine learning interactions. We note that embedding prior experience into automatic algorithm bias detection and analysis techniques is significant. The definition of fairness needs to be integrated with the laws and regulations of each country and the concept of social equity to avoid narrow technical solutions. Furthermore, the prior experience of different managers should be integrated into algorithms.

In addition, multi-domain collaborative algorithmic fairness is of significance for constructing socially responsible AI. Efforts should be made for understanding the root causes of unfairness and alleviate the cross-domain problem based on algorithm fairness. For example, the difference in loan amount between different gender groups may be considered discriminative, but it may originate from different treatments (i.e., salary) in the workplace, which may be related to the discrimination they experience at the time of enrollment. When solving the problems of discrimination in banking and recruitment seperately, different institutions may govern discrimination in terms of different fairness definitions to avoid possible losses, whereas these definitions may conflict with each another. Therefore, unified cross-disciplinary and inter-institutional algorithmic fairness techniques should be developed to build a better socio-economic ecosystem.

Another possible research direction in algorithmic fairness is to develop dynamic algorithmic fairness strategies. Current fairness studies are of limited help in real-world fairness governance because they mostly focus on passive and static fairness without considering the dynamic nature of fairness in reality. In fact, fair algorithms will influence the decision-making directions in the future applications, and consequently will affect the bias level of the subsequent input data. Therefore, it is required to dynamically adjust algorithmic fairness metrics and unfairness removal mechanisms, and to enhance the fairness of the algorithm from a long-term perspective.

7.2.3 Balancing the trade-off between performance and fairness

Building fair and reliable algorithms is the foundation of trustworthy machine learning algorithms. However, satisfying fairness notions may decrease the accuracy of the model. When protected attributes are associated with predictions, such as recidivism, it is difficult to achieve high accuracy if predictive attributes like race, poverty, unemployment, and social marginalization are excluded. There is extensive research discussing the trade-off between algorithms' performance and fairness. It is an inherent problem because the fair machine learning model is required to satisfy extra constraints: fairness metrics. In addition, as fairness notions vary with situations, it's necessary to adjust the trade-off strategy in different scenarios. Therefore, building a fair and still accurate model is a promising field in algorithmic fairness.

Author contributions
YZ had the idea for the article, XW and YZ performed the literature search and data analysis, and XW, RZ and YZ drafted the work. All authors read and approved the final manuscript.

Funding
This work was supported in part by the National Natural Science Foundation of China under Grant 71702066, in part by the National Social Science Fund of China under Grant 17BGL230, and in part by the Institute of Distribution Research, Dongbei University of Finance and Economics undet Grant IDR2021YB004.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that there is no conflict of interest.

Author details
[1]School of Management, Wuhan University of Technology, Wuhan 430070, China. [2]Research Institute of Digital Governance and Management Decision Innovation, Wuhan University of Technology, Wuhan 430070, China. [3]Management School, Lancaster University, Lancaster LA1 4YX, UK.

Received: 21 July 2022 Revised: 27 September 2022 Accepted: 8 October 2022
Published online: 10 November 2022

References

Adel, T., I. Valera, Z. Ghahramani, and A. Weller. 2019. One-network adversarial fairness. *Proceedings of the 3rd Conference AAAI on Artificial Intelligence* 33: 2412–2420.

Avin, C., I. Shpitser, and J. Pearl 2005. Identifiability of Path-Specific Effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, San Francisco, CA, USA, pp. 357-363. Morgan Kaufmann Publishers Inc.

Barocas, S., and A.D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (3): 671–732. https://doi.org/10.2139/ssrn.2477899.

Bellamy, R.K.E., K. Dey, M. Hind, S.C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K.N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K.R. Varshney, and Y. Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943.

Beretta, E., A. Santangelo, B. Lepri, A. Vetró, and J.C. De Martin. 2019. The invisible power of fairness. How machine learning shapes democracy. arXiv:1903.09493.

Bickel, P.J., E.A. Hammel, and J.W. O'Connell. 1975. Sex Bias in Graduate Admissions: data from Berkeley: measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science* 187 (4175): 398–404.

Brown, D.P., C. Knapp, K. Baker, and M. Kaufmann. 2016. Using Bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health Services Research* 51 (3): 1095–1108.

Calders, T., and S. Verwer. 2010. Three Naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21 (2): 277–292.

Calmon, F.P., D. Wei, K.N. Ramamurthy, and K.R. Varshney. 2017, April. Optimized Data Pre-Processing for Discrimination Prevention. arXiv:1704.03354.

Carey, A.N., and X. Wu. 2022. The Causal fairness field guide: perspectives from social and formal sciences. *Frontiers in Big Data* 5: 892837.

Caton, S. and C. Haas. 2020. Fairness in machine learning: a survey. arXiv:2010.04053.

Chen, J., N. Kallus, X. Mao, G. Svacha, and M. Udell 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, pp. 339-348. Association for Computing Machinery.

Cheng, L., K.R. Varshney, and H. Liu. 2021. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. arXiv:2101.02032.

Cheng, M., M. De-Arteaga, L. Mackey, and A.T. Kalai 2021. Are You Man Enough? Even Fair Algorithms Conform to Societal Norms. In *38th ICML Workshop on Socially Responsible Machine Learning*, pp. 7.

Cheng, L., S. Ge, and H. Liu. 2022. Toward Understanding Bias Correlations for Mitigation in NLP. arXiv:2205.12391.

Chiappa, S. 2019. Path-Specific Counterfactual Fairness. In *the 33rd AAAI Conference on Artificial Intelligence*, 7801–7808. Honolulu.

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

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq 2017, August. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax NS Canada, pp. 797–806. ACM.

d'Alessandro, B., C. O'Neil, and T. LaGatta. 2017. Conscientious classification: a data scientist's guide to discrimination-aware classification. *Big Data* 5 (2): 120–134.

Danks, D. and A.J. London 2017. Algorithmic Bias in Autonomous Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp. 4691–4697.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York, NY, USA, pp. 214-226. Association for Computing Machinery.
Editorial. 2016. September. *More accountability for big-data algorithms. Nature* 537 (7621): 449–449.
Elliott, M.N., P.A. Morrison, A. Fremont, D.F. McCaffrey, P. Pantoja, and N. Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9 (2): 69–83.
Feldman, M., S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney NSW Australia, pp. 259–268. ACM.
Galles, D. and Pearl, J. 2013. Testing Identifiability of Causal Effects. arXiv:1302.4948.
Garg, S., V. Perot, N. Limtiaco, A. Taly, E.H. Chi, and A. Beutel. 2019. Counterfactual fairness in text classification through robustness.
Gelfand, M.J., M. Higgins, L.H. Nishii, J.L. Raver, A. Dominguez, F. Murakami, S. Yamaguchi, and M. Toyama. 2002. Culture and egocentric perceptions of fairness in conflict and negotiation. *Journal of Applied Psychology* 87 (5): 833–845.
Hardt, M., E. Price, and N. Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.
Hasnain-Wynia, R., D.M. Weber, J.C. Yonek, J. Pumarino, and J.N. Mittler. 2012. Community-level interventions to collect race/ethnicity and language data to reduce disparities. *The American Journal of Managed Care* 18 (6 Suppl): s141-147.
Huan, W., Y. Wu, L. Zhang, and X. Wu 2020. *Fairness through Equality of Effort*, pp. 743–751. New York, NY, USA: Association for Computing Machinery.
Huang, W., Y. Wu, and X. Wu. 2020. Multi-cause discrimination analysis using potential outcomes. In *Social, Cultural, and Behavioral Modeling*, ed. R. Thomson, H. Bisgin, C. Dancy, A. Hyder, and M. Hussain, 224–234. Cham: Springer International Publishing.
Judea, P. 2009. *Causality: Models*. Reasoning and Inference: Cambridge University Press.
Jung, J., C. Concannon, R. Shroff, S. Goel, and D.G. Goldstein. 2017. Simple rules for complex decisions. arXiv:1702.04690.
Kallus, N., X. Mao, and A. Zhou. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* 68 (3): 1959–1981.
Kamiran, F. and Calders, T. 2009. Classifying without discriminating. In *Proceedings of the 2nd International Conference on Computer, Control and Communication*, Karachi, Pakistan, pp. 1–6. IEEE.
Kamiran, F., and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (1): 1–33.
Kamiran, F., T. Calders, and M. Pechenizkiy 2010. Discrimination Aware Decision Tree Learning. In *Proceedings of 2010 IEEE International Conference on Data Mining*, Sydney, Australia, pp. 869–874. IEEE.
Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer, In *Machine Learning and Knowledge Discovery in Databases*, eds. Hutchison, D., T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, P.A. Flach, T. De Bie, and N. Cristianini, Volume 7524, 35–50. Berlin, Heidelberg: Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science.
Khademi, A., S. Lee, D. Foley, and V. Honavar 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, New York, NY, USA, pp. 2907–2914. Association for Computing Machinery.
Kilbertus, N., M. Gomez-Rodriguez, B. Schölkopf, K. Muandet, and I. Valera. 2020. Fair Decisions Despite Imperfect Predictions. arXiv:1902.02979.
Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807.
Kusner, M.J., J.R. Loftus, C. Russell, and R. Silva. 2018. Counterfactual Fairness. arXiv:1703.06856.
Lambrecht, A., and C.E. Tucker. 2019. Algorithmic Bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Science* 65 (7): 2947–3448.
Lepri, B., N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. 2018. December fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. *Philosophy & Technology* 31 (4): 611–627.
Loftus, J.R., C. Russell, M.J. Kusner, and R. Silva. 2018. Causal Reasoning for Algorithmic Fairness. arXiv:1805.05859.
Lum, K. and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. arXiv:1610.08077.
Luong, B.T., S. Ruggieri, and F. Turini 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '11*, San Diego, California, USA, pp. 502. ACM Press.
Makhlouf, K., S. Zhioua, and C. Palamidessi. 2022. Survey on Causal-based Machine Learning Fairness Notions. arXiv:2010.09553.
Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54 (6): 1–35.
Nabi, R. and I. Shpitser 2018. Fair inference on outcomes. In *Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence*, pp. 1931–1940. AAAI.
Niu, Y., K. Tang, H. Zhang, Z. Lu, X.S. Hua, and J.R. Wen 2021. Counterfactual VQA: A Cause-Effect Look at Language Bias. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 12695–12705. IEEE.
Pearl, J. 2009. Causal inference in statistics: an overview. *Statistics Surveys* 3: 96–146.
Pearl, J. 2012. The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science* 13 (4): 426–436.
Pearl, J. 2012. The mediation formula: a guide to the assessment of causal pathways in nonlinear models. In *Wiley series in probability and statistics*, 1st ed., ed. C. Berzuini, P. Dawid, and L. Bernardinelli, 151–179. Wiley.
Pearl, J., and D. Mackenzie. 2018. *The book of why: the new science of cause and effect*. UK: Allen Lane.
Pearl, J., M. Glymour, and N.P. Jewell. 2016. *Causal inference in statistics: a primer*. Chichester: John Wiley & Sons Ltd.
Pessach, D., and E. Shmueli. 2023. A review on fairness in machine learning. *ACM Computing Surveys* 55 (3): 1–44.
Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K.Q. Weinberger. 2017. On Fairness and Calibration. arXiv:1709.02012.
Saxena, N.A., K. Huang, E. DeFilippis, G. Radanovic, D.C. Parkes, and Y. 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*, Honolulu HI USA, pp. 99–106. ACM.
Shpitser, I. and Pearl, J. 2007. What Counterfactuals Can Be Tested. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, Vancouver, BC, Canada, pp. 352–359.
Shpitser, I. and J. Pearl. 2012. Identification of Conditional Interventional Distributions. arXiv:1206.6876.
Teodorescu, M., L. Morse, Y. Awwad, and G. Kane. 2021. Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quarterly* 45 (3): 1483–1500.
Verma, S. 2019. Weapons of math destruction how big data increases: inequality and threatens democracy. *Vikalpa: The Journal for Decision Makers* 44 (2): 97–98.
Žliobaite, I., F. Kamiran, and T. Calders 2011. Handling Conditional Discrimination. In *Proceedings of the 11th International Conference on Data Mining*, Vancouver, BC, Canada, pp. 992–1001. IEEE.
Wu, Y., L. Zhang, and X. Wu 2018. On Discrimination Discovery and Removal in Ranked Data using Causal Graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom, pp. 2536–2544. ACM.
Wu, Y., L. Zhang, and X. Wu 2019a. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, pp. 1438–1444.
Wu, Y., L. Zhang, X. Wu, and H. Tong 2019b. PC-Fairness: a unied framework for measuring causality-based fairness. In *PRoceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, Volume 32, Vancouver.

Xu, R., P. Cui, K.Kuang, B. Li, L. Zhou, Z. Shen, and W. Cui 2020. Algorithmic Decision Making with Conditional Fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event CA USA, pp. 2125–2135. ACM.

Yang, Z.K. and J. Feng 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Volume 34, pp. 9434–9441. AAAI.

Zafar, M.B., I. Valera, M.G. Rodriguez, and K.P. Gummadi 2017a. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. arXiv:1610.08452.

Zafar, M.B., I. Valera, M.G. Rodriguez, and K.P. Gummadi. 2017b. Fairness constraints: mechanisms for fair classification. arXiv:1507.05259.

Zhang, Y. 2018. Assessing fair lending risks using race/ethnicity proxies. *Management Science* 64 (1): 178–197.

Zhang, L., Y. Wu, and X. Wu. 2016a. On Discrimination Discovery Using Causal Networks, In *Social, Cultural, and Behavioral Modeling*, eds. Xu, K.S., D. Reitter, D. Lee, and N. Osgood, Volume 9708, 83–93. Cham: Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Zhang, L., Y. Wu, and X. Wu 2016b. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 2718–2724.

Zhang, L., Y. Wu, and X. Wu 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp. 3929–3935.

Zhang, L., Y. Wu, and X. Wu. 2019. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering* 31 (11): 2035–2050.

Publisher's Note